CIMAT

Centro de Investigación en Matemáticas, A.C.

Del Registro Fósil a la Tasa de Extinción

Tesis

Que para obtener el grado de:

Maestro en Ciencias con Especialidad en Probabilidad y Estadística

Presenta

Pedro Orozco del Pino

Director de tesis:

Miguel Nakamura Savoy

Guanajuato, Guanajuato. Diciembre 2016

Mariana, eres mi primer motor. Gracias por tu apoyo incondicional, tus inmesurables sacrificios y por creer en mí en todo momento. Sin lugar a dudas este es un triunfo de los dos y me siento muy afortunado de tenerte en mi vida. Este trabajo es para ti y un poco para el pequeño monstruito que nos acompañó el segundo año. Los dos son mi inspiración y dueños de mi trabajo y de mi vida.

Agradecimientos

Este debería ser el capítulo más largo pero por mi falta de habilidad literaria no puedo expresar cabalmente lo agradecido que me siento con todos.

A mis papás Alberto y María Emilia quienes fueron actores importantísimos durante estos dos años. Sin sus consejos y su apoyo yo no habría podido concluir con éxito esta etapa. Sin restarle importancia le quiero agradecer a mi suegris Martha, quien también se mostró siempre dispuesta a darme su apoyo en todo momento. Mi cariño y admiración sólo crece cada día hacia ustedes. Muchas gracias.

Mis dos tutores Eloísa y Miguel a quienes tengo mucho que agradecerles por ser más que tutores. Siempre dispuestos a oír mis inquietudes y ayudarme a tomar las decisiones académicas difíciles. También quisiera reconocer al Profesor Pérez-Abreu, quien me enseñó que podemos aprender de cualquiera y cuyas clases siempre fueron una exquisitez. Lamento Profesor que nunca le he podido hablar de tú como me lo pidió en las primeras clases. Los tres siempre mostraron su mejor disposición para ayudarme en mi carrera profesional y académica. Para mí son un ejemplo de auténticos educadores y mentores. Por último, quiero agradecer a todos los profesores que tuve en estos dos años. Su enseñanzas me han hecho crecer profesional y académicamente más de lo que jamás habría esperado.

En especial, agradezco al Dr. Pablo del Monte Luna por la recolección de los datos utilizados en esta tesis y que amablemente compartió con nosotros. Del mismo modo, por el planteamiento del problema en el que se basa este trabajo y por la discusión sobre el tema, que en su carácter de especialista, aportó para el análisis del mismo. También quiero agradecer al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo otorgado para la realización de los estudios de la Maestría en Ciencias con especialidad en Probabilidad y Estadística que realicé en el Centro de Investigación en Matemáticas A. C. (CIMAT), a través de la beca 634806/572797.

A mis compañeros les debo mucho. Estoy muy agradecido por el grupo de amigos que formé durante estos dos años. Siempre me hicieron sentir en casa, sobretodo en los cumpleaños cuando mágicamente aparecían pasteles en los cubículos. Especialmente le quiero agradecer a Desastritos alias Desy; con su ejemplo aprendí a siempre esperar lo mejor de las situaciones y quien cuidó a Sebastián muchas veces con mucho cariño. Al equipo de futbol con el peor nombre del mundo y a su porra incondicional (especialmente Chio y Caro) les agradezco ser parte de la distracción más importante. A Cricelio le agradezco su humor y el haberme robado tantos goles; sé que lo hiciste por el bien del equipo. A Tulio y Sharo les agradezco habernos recibido con tan especial calidez en su hogar tantas veces, por ser nuestros amigos desde el primer momento (de Mariana, Sebastián y míos) y por su ejemplo de familia.

A mis hermanos Beto, Emi y Eduaro. Gracias por la complicidad, por ser tan distintos y por ser tan extraordinarios hermanos. Mis abuelitas, tías, tíos, primos, primas y sobrinos. Tener una familia tan grande y tan unida me da la

fuerza para querer ser mejor y así ser merecedor de tanta fortuna.

Al final lo más importante. Una vez más quiero agradecer a Mariana. Es aquí donde las palabras no alcanzan, pero espero que mi trabajo demuestre cuanto te agradezco que estés conmigo. Eres tú, lo fuiste y lo serás siempre.

Prefacio

La extinción de especies es un proceso natural que constantemente va cambiando con el tiempo pues depende de factores como cambios climáticos, desastres naturales e interacciones complejas entre especies existentes. El registro fósil es una manera de observar hacia el pasado y obtener información acerca del tiempo en que una especie existió y por lo tanto también de la extinción. Sin embargo, el registro fósil no es un reflejo directo de la evolución de las especies debido a que hay factores externos que afectan al proceso de fosilización y que la forma de registrar los fósiles de facto constituye muestreo por encuentro, es decir que se induce sesgo debido a observación por parte del hombre. A estos factores se les llama factores de confusión, en el sentido de que enmascaran el verdadero objeto de interés: la tasa de extinción biológica.

Tras un análisis del contexto en el cual se producen datos del registro fósil, este trabajo identifica los elementos primordiales que debe contener un modelo estadístico para la distribución del tiempo de existencia de una especie, que incorpore a los factores de confusión. Se proponen caracterizaciones matemáticas de estos elementos, para producir una familia general de distribuciones para datos observados en el registro fósil, y conteniendo parámetros para inferir acerca de la tasa de extinción. Se revisa la propuesta general del modelo estadístico y se exploran tres propuestas específicas mediante técnicas de simulación.

Índice general

Capítulo 1

Contexto

1.1. La Extinción y sus componentes

Al igual que para cualquier individuo de cualquier especie la muerte es inevitable, la desaparición total de todos los individuos de una especie a partir de algún momento y en adelante es el destino ineludible de cualquier especie. Esto quiere decir que asegurar que todas las especies en algún momento se extinguirán es tan difícil de negar como la mortalidad de los seres vivos. De hecho se estima que el 99.9 % de las especies que han existido ya están extintas [5]. La desaparición total de la vida en el planeta no ocurre debido a la radiación adaptativa y la evolución, que son los procesos por los que surgen nuevas especies. Para ser más precisos la radiación adaptativa es el proceso de especiación por el cual una o varias especies llenan nichos ecológicos [2] y la evolución es el cambio en herencia genética fenotípica de las poblaciones biológicas a través de las generaciones. Los datos que se utilizan en este trabajo no son observaciones directas de la interacción de estos procesos por lo que se identificarán los elementos estadísticos que toman en cuenta los componentes que afectan la información que se reflejan en nuestros datos acerca de la extinción.

En Huang et al. [8] se conceptualiza a la evolución mediante estructuras arboladas lo cual se aprovechará para utilizar procesos de ramificación en el modelo que propondremos. En estas estructuras el tronco representa la especie "madre", cada rama representa las distintas especies que surgieron a partir de la especie "madre", el final e inicio de una rama representan la extinción y surgimiento de la especie respectivamente y la longitud de las ramas representa el tiempo que una especie tardó en extinguirse. De esta manera el tiempo de extinción de una especie será considerado como un tiempo de vida con el objetivo de utilizar herramientas de análisis de supervivencia de una forma natural al momento de definir el objeto estadístico que deseamos estimar. Más adelante, en el capítulo de modelación, se va a profundizar más acerca de los componentes del modelo general que se va a proponer.

Observemos que a distintas alturas del árbol (ver Figura 1.1), a lo que nos referiremos como un tiempos geológicos, el número de especies puede ser diferente. Por ejemplo, en color morado tenemos un tiempo geológico con cinco especies y en color azul un tiempo geológico con diez especies, a esta cantidad de especies se denomina riqueza biológica. La riqueza biológica es necesaria para definir una medición de la extinción, la cual se denomina tasa de extinción y se puede calcular de dos maneras. La primera es la cantidad total de especies extintas en un tiempo geológico, lo cual

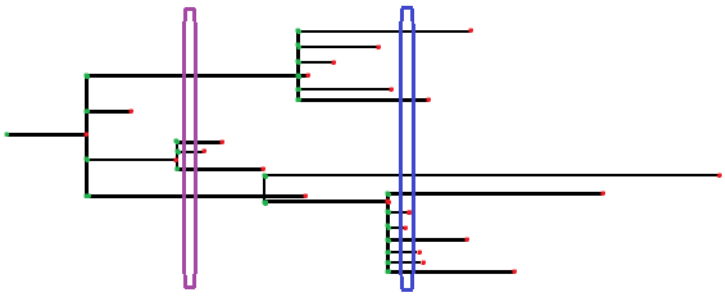

Figura 1.1: Estructura de árbol para visualizar la evolución. Se señala en color morado un tiempo geológico en el que la especie madre ya evolucionó en 5 distintas especies y el color azul señala un tiempo geológico en el que la especie madre ya evolucionó en 10 especies distintas. Los puntos verdes representan el inicio de una especie y el punto rojo sobre la línea horizontal que surge del punto verde representa la extinción de la especie, de esta manera la longitud de la línea será el tiempo que existió la especie.

resultaría en una tasa de extinción de cero para el tiempo geológico morado y de dos para el tiempo geológico azul. La segunda es el cociente expresado en porcentaje entre el número de especies extintas y la riqueza biológica, es decir una tasa de extinción de 0 % para el tiempo geológico morado y 20 % para el tiempo geológico azul. Para una explicación más detallada de la tasa de extinción véase Barnosky et al. [6]. La segunda forma de medir la tasa de extinción es la que se va a utilizar en este trabajo pues permite comparar la tasa de extinción para distintos tiempos geológicos.

El cambio en la tasa de extinción a lo largo del tiempo que se ilustra en la Figura 1.1 b) es drástico en algunas ocasiones, por ejemplo debido a las extinciones masivas de especies. Por ello que no es razonable suponer que esta tasa es constante para distintos tiempos geológicos. De aquí que el comportamiento de la tasa de extinción está relacionado con cualquier factor que cambie las condiciones geológicas, atmosféricas, químicas, etcétera, del planeta. Estos factores son algunos ejemplos de los componentes que confunden la información de la extinción que se encuentra en nuestros datos.

Cualquier evento que ocurra en el planeta deja algún tipo de huella y en el caso de las especies es mediante la fosilización. Esto quiere decir que los fósiles son una manifestación de la extinción y por lo tanto también de la tasa de extinción. Al conjunto de fósiles fechados se va denominar registro fósil y la base de datos con la se va a trabajar es un listado de especies en la que para cada una tenemos la fecha del fósil más antiguo y la fecha del más reciente. Es muy importante mencionar que no todas las especies son capaces de fosilizar (las especies con estructuras corporales suaves como los moluscos no se fosilizan) o habitaron en un ecosistema con las condiciones climáticas y geológicas adecuadas para fosilizar por lo que el registro fósil no puede contener a todas las especies que han habitado al planeta. Por lo tanto debemos tomar en cuenta que si nuestros datos provienen de fósiles entonces no se encontrarán todas las especies que han existido. Es decir, aunque observáramos todos los fósiles que existen no tendríamos representadas a todas las especies que han existido en el planeta.

El registro fósil es en realidad una base de datos obtenida mediante la excavación de las capas geológicas. Dicha excavación no es uniforme, es decir que algunas capas geológicas se han explorado más que otras. Esto deriva en que la cantidad de fósiles encontrados en dos capas geológicas puede diferir debido a un muestreo disparejo y no

Figura 1.2: Estructura de la información para cada especie en el registro fósil.

necesariamente a una discrepancia en riqueza de fósiles. A este factor de confusión se le conoce como esfuerzo de muestreo y se va a tener en cuenta en el modelo. En Marshall [9] se resumen los enfoques con los que se ha abordado este problema. En Alroy [7] se menciona una manera de remover el esfuerzo de muestreo mediante remuestreo. A diferencia de estos trabajos, en esta tesis el esfuerzo de muestreo no va a ser el único factor de confusión que se considere en el modelo.

Cada caso que se tiene en el registro fósil representa a una única especie; en particular la base de datos que se va a analizar tiene 35,868 casos y esto quiere decir que se tiene información de 35,868 especies distintas. La base de datos fue proporcionada por el Dr. Pablo del Monte Luna quien labora en el Centro Interdisciplinario de Ciencias Marinas (CICIMAR). Esta base de datos es un compendio de toda la literatura en Paleontología hecho por Sepkoski y que es mencionado por primera vez en [4]. Para entender la estructura de la información que se tiene de cada especie nos apoyaremos en la Figura 1.2. En cada caso se tiene la fecha de inicio y de fin del estrato geológico más antiguo (intervalo de color verde) y más reciente (intervalo de color rojo) en el cual ha sido encontrado al menos un fósil de esa especie (representado por un punto azul). Esta estructura nos induce datos censurados por ambos lados ya que en realidad no sabemos la verdadera posición de los puntos azules ilustrados en la Figura 1.2. Este tipo de censura la presentan el 83 % de los datos pues son especies que ya están extintas. El resto de las especies que se encuentran en la base de datos son especies que siguen existiendo en el planeta y no se tiene la información que se ilustra con el intervalo en color rojo que muestra la Figura 1.2. No consideramos el problema de modelar la cenura en este trabajo debido a que se priorizó en favor de modelar los factores de confusión. Sin embargo se mencionará en la sección de trabajo futuro la importancia de incluir censura en el desarrollo posterior del análisis.

Como última observación acerca del registro fósil mencionaremos que el grado de resolución para cada especie es distinto, es decir que para algunos casos la información respecto a la aparición (o extinción) es un intervalo más amplio que en otros casos. Esto se debe a que las capas geológicas tienen una clasificación jerárquica, la cual tiene alrededor de once periodos geológicos y cada uno tiene subdivisiones llamadas épocas geológicas y cada una de ellas tiene a su vez subdivisiones llamadas etapas geológicas. Es decir, que algunas especies la información se encuentra a nivel de periodo, de otras a nivel de época y de otras a nivel de etapa. Por ejemplo si la especie A aparece en la base de datos con un nivel de resolución de periodo y la especie B aparece con un nivel de resolución de época, entonces la especie A tiene una resolución menor en la información y tiene más censura que la especie B. Para fines de un primer análisis vamos a ignorar esta resolución y se tomará como tiempo de vida de cada especie el mínimo del intervalo antiguo (color verde Figura 1.2) y el máximo del intervalo reciente (color rojo Figura 1.2). Para un trabajo futuro se debe tomar en cuenta la resolución de los datos al momento de incluir la censura en el análisis.

1.2. Motivación y alcance de la tesis

Esta tesis es el primer acercamiento a una inquietud planteada por el Dr. Pablo del Monte Luna. Esta surge a partir de que en Nakamura et al. [3] se estima el cambio en la tasa de extinción de especies marinas con datos de últimos avistamientos de hasta 500 años de antigüedad y se propone una cuantificación de la tasa de extinción para la época actual que toma en cuenta la incertidumbre que genera trabajar con el concepto de últimos avistamientos en lugar de una medición directa de la extinción. Sin embargo, es de interés saber si el aumento de la tasa de extinción desde la aparición del hombre no es por algún otro factor que coincidió con el inicio de la humanidad. Para esto es necesario comprender el comportamiento de la tasa de extinción antes de que la raza humana hiciera su aparición en la Tierra para poder hacer una comparación adecuada.

La pregunta básica es inferir acerca del comportamiento de la tasa de extinción durante los últimos 540 millones de años para hacer comparación con la presencia del hombre y sin la presencia del hombre. La tesis no va a responder dicha pregunta básica sino que va a delinear los elementos que se deben tomar en cuenta para poder resolverla.

Es importante ser más precisos respecto a los elementos de los que se va a hablar y aquellos que se dejarán de lado para trabajos futuros. Los alcances de la tesis se van a dividir en dos ejes principales: proponer una familia de modelos que sea adecuada para resolver la pregunta básica e identificar las agravantes que son inherentes al problema. En el desarrollo del trabajo quedará claro al lector que resolver la pregunta básica es un problema muy complejo y de esta manera los alcances limitados de la tesis serán evidentes^{o}.

Las agravantes de la pregunta básica tienen que ver con los componentes de confusión antes mencionados que provocan el que en el registro fósil no se manifieste de manera transparente el proceso de extinción. Lo que quiere decir esto está relacionado con dos situaciones principalmente: la primera es que intervienen condiciones naturales que evitan que todas las especies fosilicen y la segunda es que los procesos humanos que intervienen en la recolección de fósiles son imperfectos y por lo tanto presentan sesgos que deben ser tomados en cuenta. Estas situaciones, que llamaremos factores de confusión a partir de este momento, provocan una sobrevaloración de la información que contiene el registro fósil acerca de la extinción.

Los factores de confusión son muy diversos y pueden mencionarse como ejemplos: el proceso de fosilización de las especies, la censura de los datos o el muestreo por encuentro. El último se refiere a los factores naturales y humanos que intervienen en que un fósil sea encontrado en la excavaciones. Sin lugar a dudas existen otros factores de confusión que afectan el registro fósil que no se mencionan debido a que enumerarlos no es objeto de este trabajo. La inclusión de estos factores en el modelo será a través de una función que va a capturar la incertidumbre de todos los factores. Cómo se verá más adelante en las simulaciones esta función contiene una gran cantidad de incertidumbre que se reflejará en la verosimilitud obtenida con el modelo. Sin embargo, este enfoque tiene la ventaja que incluye los factores de confusión en el modelo a pesar de no tener ninguna información de los mismos. La desventaja es que no es interpretable y por lo tanto no es de utilidad para hacer inferencia respecto a los factores de confusión, no obstante cumple con el propósito de la tesis. En el capítulo de modelación se profundizará más acerca de los elementos que tiene esta función junto con una justificación basada en el trabajo interdisciplinario con el Dr. Del Monte.

La propuesta de una familia de modelos va a consistir en identificar, mas no especificar, los elementos que un modelo debe tener para poder resolver la pregunta básica tomando en cuenta los factores de confusión. Estos elementos van a representar al proceso evolutivo de las especies y los procesos, tanto naturales como humanos, que provocan que dicho

proceso deje huella en el registro fósil.

Una idea primordial de la familia de modelos que se toma en cuenta es que no hay una única especie madre que dio lugar a todas las demás sino que puede haber distintas especies madre y en diferentes tiempos geológicos. Esto da pie a pensar en que la evolución no es una única estructura arbolada sino un conjunto de muchas estructuras arboladas. Esto dará lugar a un modelo de mezclas para los tiempos de vida pues no hay información acerca de que especie madre proviene cada uno de los fósiles. Es decir, el registro fósil no tiene información acerca del número de estructuras arboladas que suponemos representan a la evolución.

Se propondrá un modelo paramétrico para facilitar la interpretación del modelo y separar los elementos que se consideran de confusión de la pregunta básica.

1.3. Plan de la Tesis

La tesis se va a desarrollar en dos capítulos: análisis exploratorio y modelación. En el análisis exploratorio se van a utilizar herramientas de visualización, exploración interactiva y estadística no paramétrica para comprender mejor la interacción del registro fósil con la extinción. La modelación va a proponer un modelo probabilístico que, basado en la información obtenida en el análisis exploratorio, tenga elementos que se puedan interpretar para responder la pregunta básica. En el capítulo de modelación simularemos escenarios distintos que sigan los lineamientos del modelo para corroborar si los datos habrían podido surgir de dicho escenario. Mediante estos dos capítulos la tesis tiene una línea de acción que consiste en entender el problema, proponer una solución y corroborar con datos. Al final estos dos capítulos dan lugar a una sección de conclusiones y trabajo futuro en la cual se mencionarán las principales aportaciones de este trabajo, sus puntos débiles y cuál es el trabajo a futuro que se sugiere.

Capítulo 2

Análisis Exploratorio

El análisis exploratorio tiene como objetivo comprender mejor la riqueza biológica, la radiación adaptativa, la evolución, la extinción misma, los tiempos de vida de las especies y la dependencia del tiempo geológico para tener sustento para los componentes del modelo. Para comprender la dependencia del tiempo geológico se va a recurrir a una herramienta interactiva para visualizar en tiempo real el cambio en las herramientas exploratorias al variar el tiempo geológico.

A pesar de que ya se ha hecho referencia es importante notar que la variable de tiempo aparece de dos maneras muy distintas y fáciles de confundir. Por un lado los datos están situados en distintos estratos geológicos, es decir que hay fósiles que pertenecen a especies que comenzaron a existir muchos millones de años antes que otras y por lo tanto las llamaremos especies "antiguas" con su contra parte de especies "jóvenes". A esta modalidad del tiempo es a lo que nos hemos referido anteriormente con tiempo geológico y continuaremos haciéndolo de esta manera. Por otro lado, también se tiene la cantidad de millones de años que una especie existió en el planeta y esto nos divide a las especies en "longevas" y "no longevas"; esta modalidad del tiempo son los tiempos de vida. Notemos que puede haber especies antiguas que no son longevas y especies jóvenes que son longevas. Un ejemplo de el primer caso sería una especie que surgió hace quinientos millones de años pero que sólo habitó el planeta durante un millón de años. Un ejemplo del segundo caso sería una especie que surgió hace sesenta millones de años y que aún no se extingue. Esta diferenciación es muy importante pues además de denotar nociones diferentes, juegan papeles muy distintos en la formulación de modelos estadísticos: en el primer caso va a participar como un índice para los parámetros del modelo y en el segundo caso constituye de facto la variable principal de interés. En pocas palabras, es de interés estudiar la distribución de los tiempos de vida como función del tiempo geológico.

2.1. Visualización del registro fósil

Para poder entender mejor cómo se manifiesta el registro fósil vamos a empezar por visualizarlo. En la Figura 2.1 se representan todos los tiempos de vida que se encuentran en nuestro registro fósil; la Figura 2.2 es un acercamiento de la Figura 2.1. En esta gráfica se puede apreciar lo siguiente:

- Las posibles extinciones masivas en las concentraciones de círculos rojos, por ejemplo al término del periodo Pérmico. Se profundizará en el concepto de extinciones masivas y su relevancia en la modelación más adelante.

- La riqueza biológica creciente al observar que cada vez hay mayor número de especies vivas. Esto se aprecia en el hecho de que hay mayor número de segmentos con una altura mayor a cien especies a partir del periodo Cretácico.

- La variación en tiempos de vida que hay en las especies. Algunas duran cientos de millones de años y otras ni siquiera se aprecia la longitud del tiempo de vida, es decir el círculo rojo está sobre el punto verde. El segundo caso es un ejemplo de especies no longevas.

- La censura por resolución del tiempo geológico. Este último aspecto se aprecia al observar que hay especies que repiten el comportamiento, es decir que aparecen y se extinguen en el mismo momento. Sin embargo, se cree que en caso de no haber censura esto último sería extremadamente inusual de que ocurriera. No obstabte, se observa frecuentemente en la gráfica, de hecho la manera de graficar el registro fósil tendría que replantearse de no tener censura.

Esta visualización del registro fósil sirvió como punto de partida para el análisis exploratorio pues no sólo ilustró los fenómenos antes mencionados sino que también hizo evidente la necesidad de incluir los factores de confusión. El ejemplo más evidente es la gran cantidad de segmentos que llegan al final de la gráfica, es decir el número de especies que según nuestro registro fósil aún no están extintas. En particular este número representa el 17 % de los datos, lo cual difiere ampliamente al 1 % que se menciona en Novacek [5].

2.1.1. Riqueza biológica creciente

El comportamiento de la riqueza biológica afecta directamente a la tasa de extinción. Por ejemplo, si la riqueza aumenta y la tasa de extinción queda constante esto implica que el número de especies que se están extinguiendo crece a la misma velocidad que la cantidad de nuevas especies. Anteriormente se mencionó que la Figura 2.1 ilustra que la riqueza biológica ha ido en crecimiento. Sin embargo, esto lo haremos más evidente con las siguientes dos gráficas. En Marshall [9] la gráfica A (Figura 2.3) representa la riqueza a través del tiempo y en la Figura 2.4 se ilustra el mismo cálculo pero hecho con nuestra base de datos. En ambos casos se aprecia que la riqueza sí tiene un comportamiento claramente creciente. Es importante mencionar que en el cálculo no hay ningún tipo de corrección por esfuerzo de muestreo u otro factor de confusión por lo que el comportamiento preciso de la riqueza biológica puede ser distinto. A pesar de la observación anterior, la tendencia a la alza es tan clara que sería muy difícil que la riqueza biológica no fuera creciente, aún si se tuviera en cuenta los factores de confusión. Con esto vamos a considerar que efectivamente la riqueza biológica ha ido en aumento a lo largo del tiempo geológico aunque no sabemos con qué magnitud.

2.1.2. Extinciones Masivas

Las extinciones masivas son uno de los factores que producen un cambio más drástico en la evolución. Usualmente se deben a fenómenos repentinos en el planeta, como por ejemplo la extinción masiva del Cretácico-Terciario que

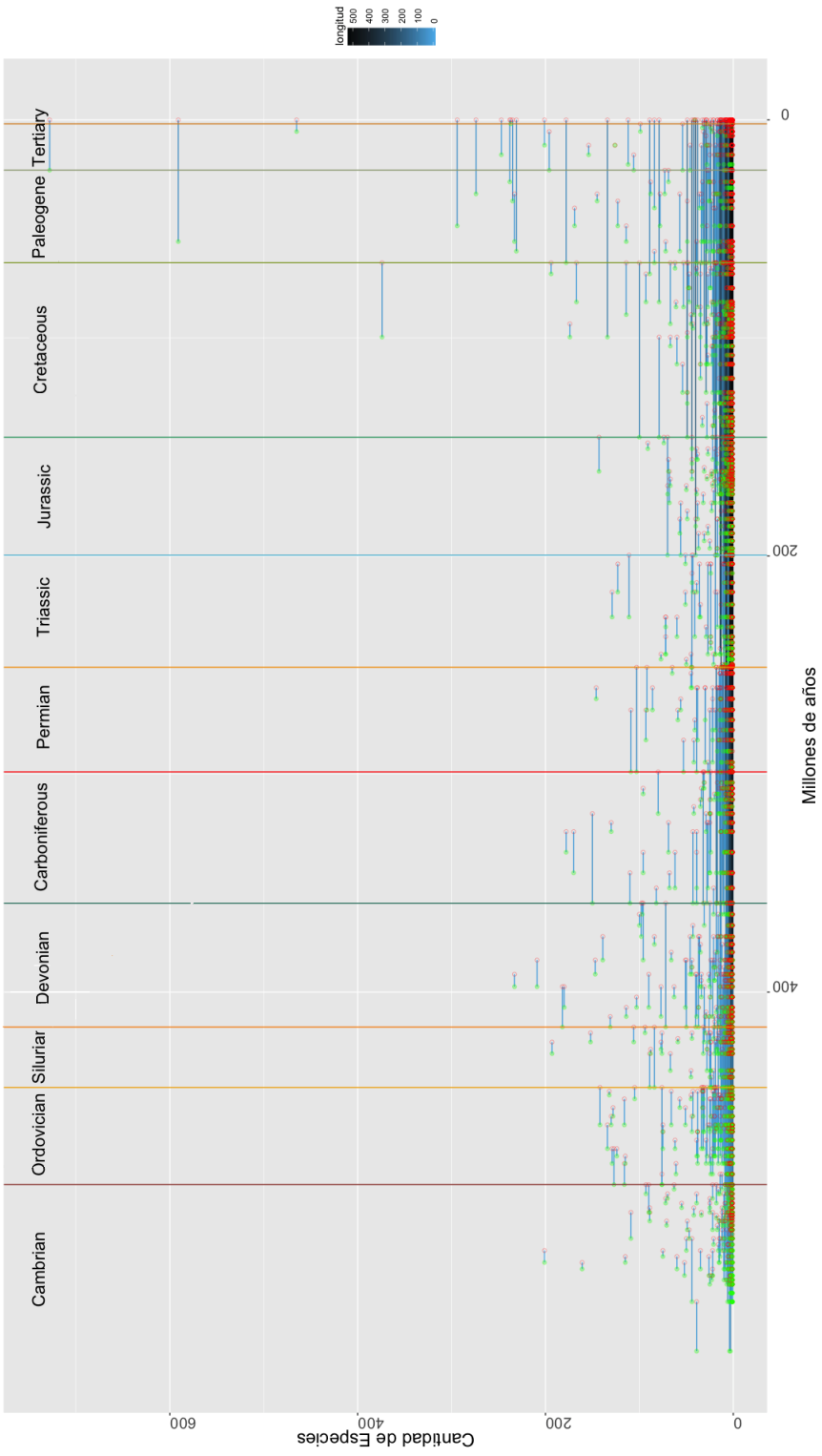

Figura 2.1: Representación del registro fósil completo. La base de datos la proporcionó Pablo del Monte y contiene 35,868 especies distintas. En la parte superior se indica el nombre de la era geológica y con líneas de distintos colores se marca el final de cada una de ellas.

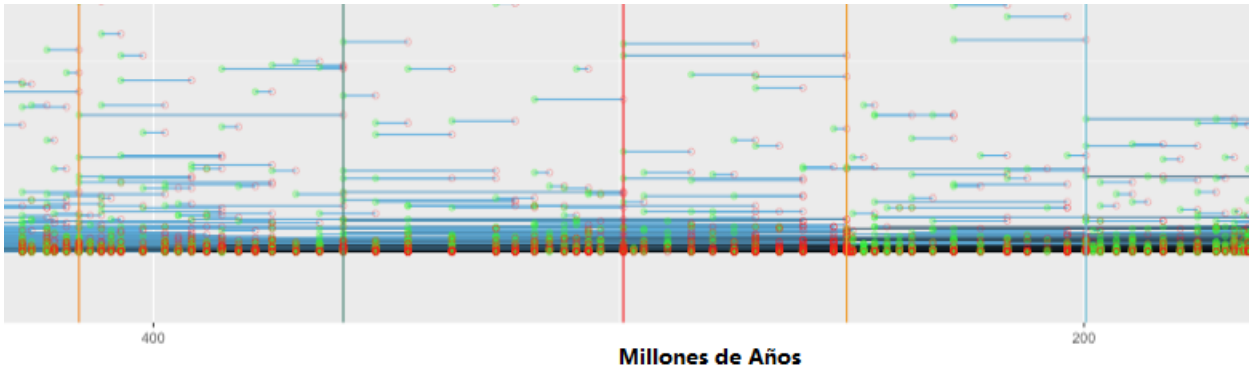

Figura 2.2: Acercamiento a la representación visual del registro fósil que se observa en la Figura 2.1. Cada segmento de línea tiene un punto verde de inicio y un círculo rojo como final. La altura de los segmentos representa la cantidad de especies que tienen ese mismo tiempo de vida durante ese mismo periodo de tiempo geológico. Los puntos verdes corresponden al surgimiento de la especie mientras que los círculos rojos se relacionan con la extinción. El color de los segementos lo determina su longitud: mientras más azules claro son más cortos y se van oscureciendo conforme el tiempo de vida que representan es mayor.

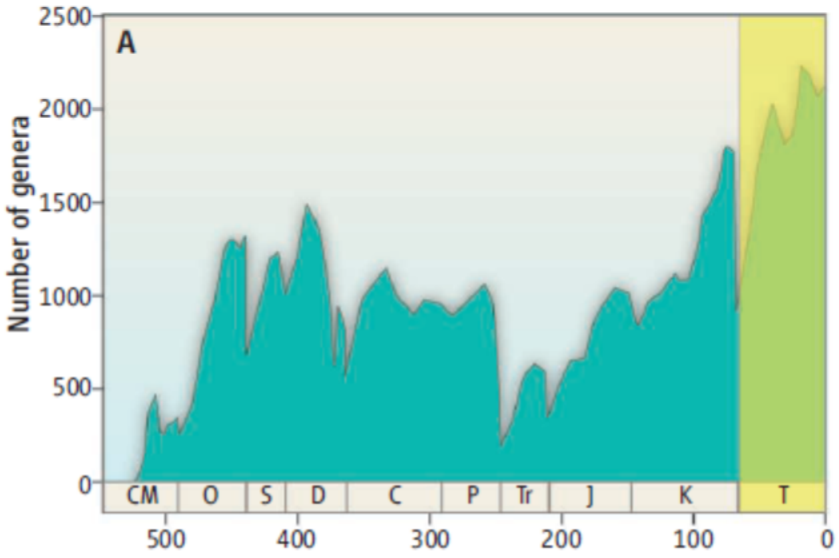

Figura 2.3: Cáculo de la riqueza biológica según Sepkoski en [4]. Esta imágen se encuentra en Marshall [9]

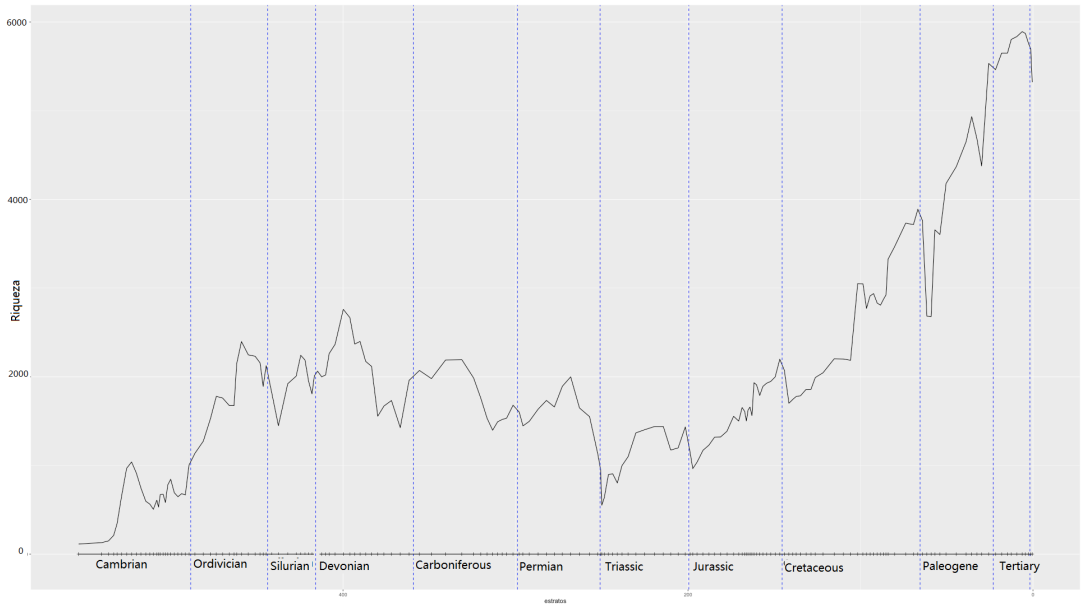

Figura 2.4: Cálculo de la riqueza biológica siguiendo el razonamiento de Sepkoski mencionado en [9] utilizando nuestra base de datos.

eliminó por completo a los dinosaurios cuando un meteorito cambió abruptamente las condiciones climáticas. Como se mencionó anteriormente el registro fósil tiene información acerca de estas extinciones. Primero describiremos de forma muy breve las cinco extinciones masivas que se conocen y que están señaladas en la Figura 2.5. La información se obtuvo de [16].

I. Ordovicio Silúrico. Esta extinción afectó principalmente a las especies marinas pues la mayoría de la vida aún habitaba en los océanos.

II. Devoniano tardío. Tres cuartas partes de las especies se extinguieron durante esta extinción masiva, sin embargo se cree que fueron una serie de extinciones a lo largo de millones de años. Es por esto que se señalan varios picos para este caso.

III. Pérmico. Se cree que esta es la extinción masiva más fuerte ya que acabo con el 96 % de las especies. Esto quiere decir que la especies actuales provienen del 4 % restante. Notemos que aunque la tasa correspondiente en la Figura 2.5 no es del 96 % si es la tasa relativa más elevada según la base de datos.

IV. Triásico Jurásico. Durante los últimos 18 millones de años del periodo Triásico existieron dos o tres fases de extinción que ocasionaron esta extinción masiva. El cambio climático, fuerte actividad volcánica y el impacto de un asteroide son las causas de esta extinción.

V. Cretácico Terciario. Esta es la extinción masiva en la cual los dinosaurios desaparecieron, sin embargo muchos otros organismos también se extinguieron en este periodo.

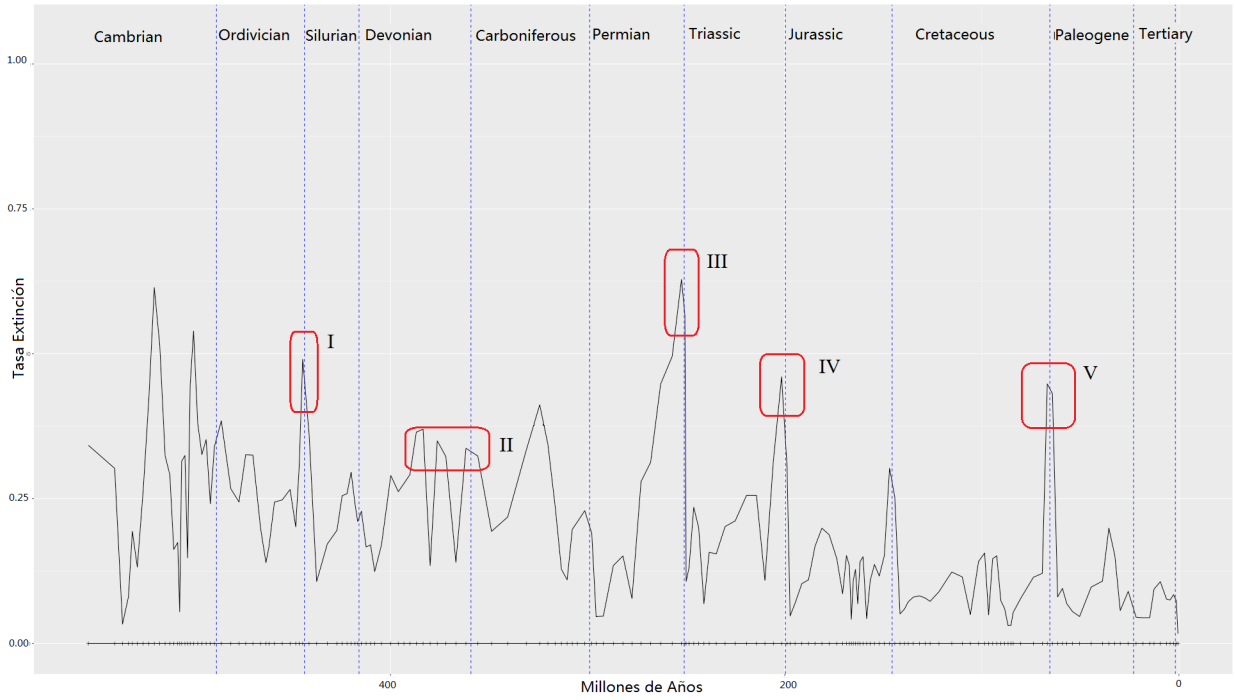

Figura 2.5: Cálculo de la tasa de extinción relativa a través del tiempo geológico. Las extinciones masivas conocidas están señaladas con rectángulos rojos y tienen el número que les corresponde en la explicación.

Una forma de saber si el registro fósil usado en esta tesis contiene información de extinciones masivas es si refleja de alguna manera estas cinco extinsiones masivas conocidas. En la Figura 2.5 se presenta la tasa de extinción relativa a través del tiempo geológico. Podemos notar que la base de datos sí contiene información de las extinciones masivas que son conocidas. Notemos que no todos los picos de la tasa de extinción están marcados como extinciones masivas conocidas y eso es parte del problema que se quiere abordar, pues no sabemos si esos aumentos de tasa de extinción son en verdad extinciones masivas o el resultado de algún factor de confusión. Esto quiere decir que el registro fósil que estamos usando contiene información de las extinciones masivas conocidas y quizá también información de extinciones masivas no identificadas en la literatura. Mientras no haya una manera de valorizar la incertidumbre asociada con la magnitud de esos picos, cualquier aseveración será aventurada o polémica. El punto de la modelación estadística que se aborda en esta tesis es comenzar a abordar la pregunta con razonamientos estadísticos formales.

2.2. Entendimiento de la relación entre tiempo geológico y tiempos de vida

Ignorar al tiempo geológico como un factor fundamental del problema sería una simplificación ingenua y poco útil pues el hecho de que el planeta Tierra ha cambiado de manera drástica durante los últimos 540 millones de años es indiscutible. Sin embargo, es importante entender cómo se han plasmado estos cambios en el registro fósil. Esto significa que queremos entender la manera en que el riesgo de extinción, la distribución del tiempo de vida y los eventos catastróficos se reflejan en el registro fósil.

Lo primero que podemos explorar es si la probabilidad de extinguirse, o de que surja una nueva especie, en tiempos geológicos pequeños es uniforme según el registro fósil. Observemos la Figura 2.6, donde se muestran dos estimaciones de densidades por Kernel mediante la función *density* de R, la cual utiliza un ancho de banda calculado por validación cruzada de manera predeterminada y es el que se utilizó en las gráficas. La de color verde es la densidad que se obtiene con la fecha del registro más antiguo de cada especie, es decir que es una aproximación del comportamiento de la radiación adaptativa. Con color rojo se ilustra la densidad estimada que se obtiene con la fecha del registro más actual de cada especie, es decir que es una aproximación (ignorando a los factores de confusión) de la extinción. Se puede notar que existe una alta heterogeneidad a través del tiempo geológico, que está representado en el eje x de la densidad, así como la presencia de valles y picos en las densidades. Estos valles y picos pueden ser el reflejo de los eventos catastróficos, extinciones masivas, cambios climáticos, etcétera. Resumiendo, la Figura 2.6 nos confirma que es muy importante considerar que el comportamiento de la extinción es variable con relación al tiempo geológico.

2.2.1. Herramienta interactiva de exploración de datos

Considerar el tiempo geológico como un elemento primordial en el modelo complica el análisis exploratorio ya que la visualización tendría que ser en tres dimensiones. Es decir que cuando construimos una herramienta gráfica para un tiempo geológico fijo que nos permita, por ejemplo, explorar la posibilidad de exponencialidad en los tiempos de vida tendremos que repetir el proceso cada vez que cambiemos el tiempo geológico. Debido a que nos interesa el comportamiento a través del tiempo geológico tendríamos que observar simultáneamente un gran número de gráficas.

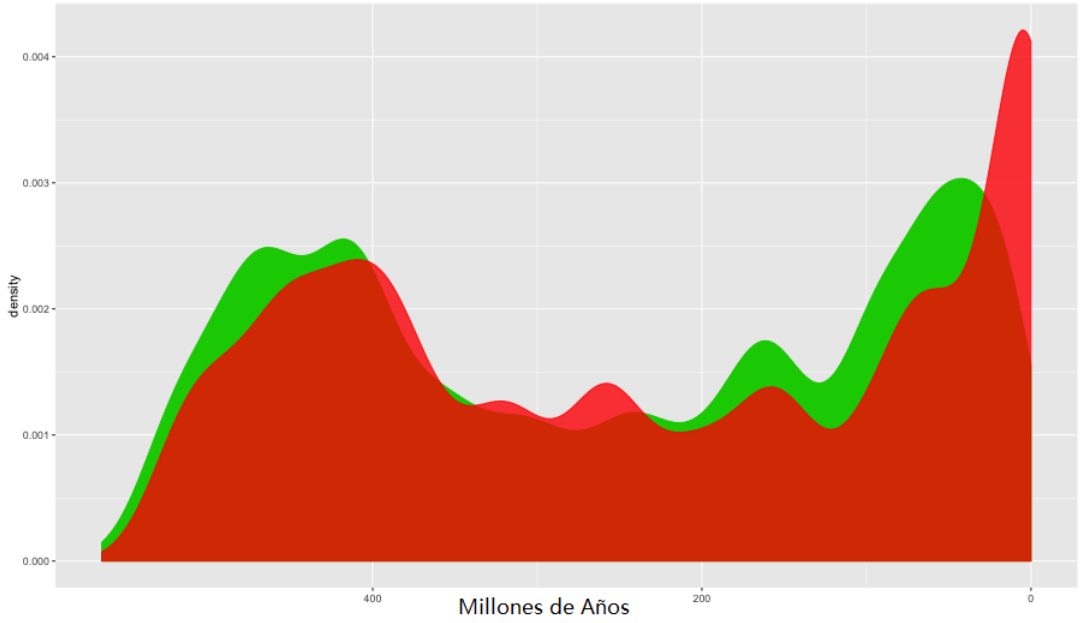

Figura 2.6: Densidades por método de Kernel de los nacimientos y extinciones según el registro fósil. Recordemos que en realidad el registro fósil contiene información censurada de la extinción y la radiación adaptativa de las especies.

De esta problemática surgió la necesidad de una herramienta que nos permitiera navegar a través del tiempo geológico de manera eficiente. A continuación vamos a describir dicha herramienta.

Primero, especifiquemos cómo se seleccionan los datos con los que cada tiempo geológico es explorado. Para un tiempo geológico fijo t se eligen los datos del registro fósil cuyos tiempos de vida lo contengan. Para entender mejor la selección de los datos recordemos que el registro fósil nos indica el intervalo de tiempo geológico en el cual han sido encontrados fósiles de cada especie, entonces para el tiempo geológico t seleccionamos los tiempos de vida de las especies cuyo intervalo contiene a t. Es decir que estamos tomando los tiempos de vida de las especies que existieron durante el tiempo geológico t.

Seleccionar los datos de esta forma nos va a permitir recorrer el tiempo geológico de manera continua y obtener los tiempos de vida de las especies que existieron en cualquier tiempo geológico además de que la visualización será posible con una herramienta interactiva. La herramienta interactiva que utilizamos se llama Shiny, que es construida por RStudio, y gracias a ella podemos programar tableros que se visualizan en cualquier navegador, tablet o teléfono inteligente y que permite que las gráficas exploratorias se actualicen en tiempo real conforme navegamos a través del tiempo geológico. En la Figura 2.7 se muestra un ejemplo de la herramienta para el caso de un tablero que grafica un estimador no paramétrico de la función de riesgo que será discutido más adelante. En esta imagen se puede apreciar un selector en la parte izquierda que sirve para seleccionar los datos con los que se construye la gráfica de la derecha. Los tableros que se construyen quedan guardados en la nube y se pueden consultar siguiendo un url. La Figura 2.7 se puede visualizar siguiendo el siguiente url https://orozcopedro.shinyapps.io/ContornosRiesgo/. De esta manera ahora podemos explorar cualquier cualidad del registro fósil que dependa del tiempo geológico con un sólo tablero en lugar

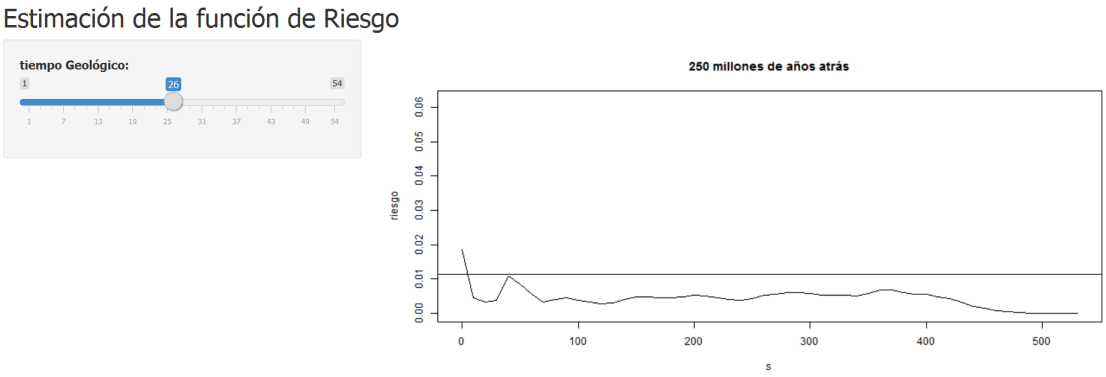

Figura 2.7: La aplicación de Shiny para visualizar al estimador no paramétrico de la función de riesgo para un tiempo geológico fijo t. En el lado izquierdo se tiene un selector del tiempo geológico y del lado derecho se ve la gráfica del estimador de la función de riesgo según Bouezmarni [15]

de una gran cantidad de gráficas.

Estimación no paramétrica de la función de riesgo

El trabajo de Nakamura et al [3] es un ejemplo de explorar un tiempo geológico fijo. En particular abordan el tiempo geológico actual. En este artículo utilizan el concepto de función de riesgo(también llamada función de hazard) de análisis de supervivencia y teoría de confiabilidad para entender mejor el riesgo de extinguirse que tiene una especie. Para esto definiremos a la función de riesgo. Si S es la variable aleatoria que representa el tiempo de vida de una especie, $F(s)$ es su distribución y $f(s)$ su densidad entonces la función de riesgo se define como sigue:

$$h(s) = \frac{f(s)}{1 - F(s)}, \tag{2.1}$$

donde s se refiere al tiempo de vida y no al tiempo geológico. La expresión 2.1 se interpreta como la tasa de extinción inmediata de una especie. Es decir que para intervalos pequeños de s, denotados Δs, la probabilidad de que una especie se extinga durante las próximas Δs unidades de tiempo es $h(s)\Delta s$.

Observemos que en nuestros caso $h(s)$ va a depender del tiempo geológico, y por lo tanto vamos a denotar con un subíndice t para indicar que se refiere a un tiempo geológico específico, es decir $h_t(s)$. Entonces interpretamos a $h_t(s)$ como la tasa de extinción inmediata que tenían las especies con s millones de años de existir y que habitaban el planeta durante el tiempo geológico t.

Para estimar a $h_t(s)$ se utilizó la propuesta de Bouezmarni [15] que utiliza un estimador con Kernel Gama con ancho de banda dependiente de los datos para eliminar el efecto frontera que usualmente los estimadores por Kernel presentan. Debido a que para proponer una distribución de los tiempos de vida vamos a basarnos en la forma que presente la gráfica del estimador no paramétrico de la función hazard, entonces es importante eliminar este efecto frontera. De lo contrario podríamos caer en el error de fundamentar la distribución de los tiempos de vida en un efecto inherente al instrumento de estimación y no en información que proviene de los datos que se esté manifiestando en la forma de la

gráfica del estimador. Se puso especial atención a que si la forma de la estimación de $h_t(s)$ fuera aproximadamente constante pues en ese caso la distribución de S sería exponencial. Esto último es conveniente pues permitiría usar procesos de ramificación con pérdida de memoria para conceptualizar la evolución además de la conveniencia analítica de la distribución exponencial.

El estimador que se propone en Bouezmarni [15] es el siguiente:

$$\hat{h}_t(s) = \frac{\hat{f}_t(s)}{1 - \hat{F}_t(s)}, \tag{2.2}$$

donde $\hat{F}_t(s)$ es el estimador de Kaplan Meier de la función de distribución de S y $\hat{f}_t(s)$ es el estimador por Kernel con Kernel Gama que se describe en Bouezmarni [15]. El efecto frontera se elimina al no asignarse peso fuera del soporte de los datos mediante la dependencia del ancho de banda con los datos.

En la Figura 2.8 se muestran 6 impresiones del tablero que se construyó para evaluar la estimación de la función de riesgo a través del tiempo geológico. En esta imagen se observa que la función de riesgo tiene cierta similitud con una función constante en todos los casos excepto en el caso (f) en el que la forma correspondería a una distribución Gama por su comportamiento decreciente de forma exponencial. Esto nos dice que es razonable que la distribución de los tiempos de vida sea similar a la distribución exponencial, hecho que vale la pena ser tomado en cuenta en la modelación. Sin embargo, este hecho no es concluyente pues se observa un comportamiento ligeramente decreciente, especialmente en los casos (a), (c), y (e) donde se alcanza a ver un pico cerca del cero. La posible exponencialidad de los datos va a ser un punto central en el capítulo de modelación. En este se va a especificar una posible explicación al comportamiento observado sin tener que descartar que los tiempos de vida tengan una distribución exponencial.

Multimodalidad de la distribución del tiempo de vida

Un modelo paramétrico para los tiempos de vida permitiría hacer inferencia interpretable acerca de la tasa de extinción a través del concepto de la función de riesgo. Para identificar características que deba tener el modelo paramétrico observaremos las características que estimaciones no paramétricas resalten de los datos. En la Figura 2.9 se muestran seis pantallas de un tablero que se construyó para visualizar el estimador por Kernel de la densidad de los tiempos de vida. Podemos apreciar que para todos los casos la mayoría de los datos están concentrados cerca del cero, es decir que las especies no longevas son más comunes que las longevas. Esta característica nos dice que un modelo exponencial podría ser factible. Sin embargo también se puede ver que existe una posible multimodalidad (casos (a), (b), (c) y (e)) y por lo tanto el modelo exponencial no sería adecuado. Este tablero nos dice entonces que debemos considerar una distribución de los tiempos que permita multimodalidad y que tenga un comportamiento similar al exponencial.

Notemos que si los tiempos de vida fueran exponenciales entonces la multimodalidad no se puede explicar ni siquiera con un modelo de mezclas. Esto sucede ya que la distribución exponencial es unimodal y con la moda en el cero y por lo tanto la única manera de crear multimodalidad sería usando exponenciales truncadas. Sin embargo, esto implicaría estimar parámetros de umbral lo cual induciría una verosimilitud con singularidades y parece ser una complicación innecesaria. Se va a mostrar en el capítulo de modelación que la multimodalidad es posible con tiempos exponenciales al introducir el concepto de probabilidad de encuentro. Otra manera de pensar esto es la siguiente. Si los tiempos de

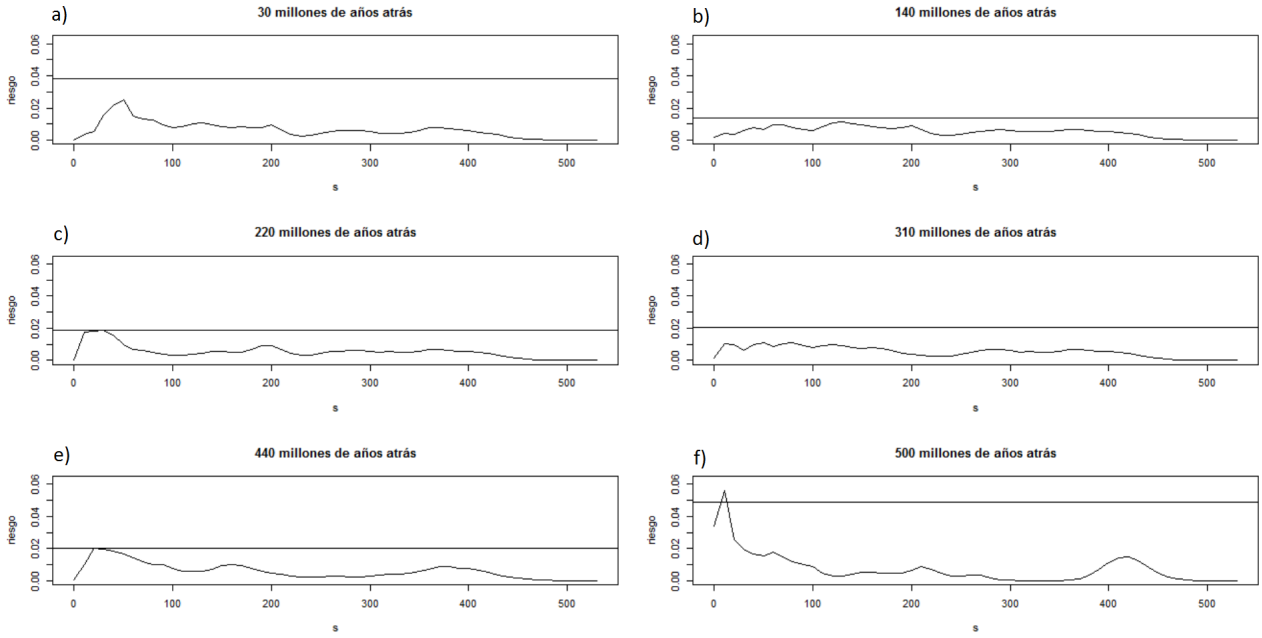

Figura 2.8: Ejemplo de $\hat{h}_t(s)$ para seis distintos tiempos geológicos obtenidos de la aplicación de Shiny https://orozcopedro.shinyapps.io/ContornosRiesgo/. Las líneas horizontales representan la función hazard asumiendo que los datos tienen una distribución exponencial. Observemos que este valor varía según la era geológica, mostrando que la tasa de extinción depende de la era geológica.

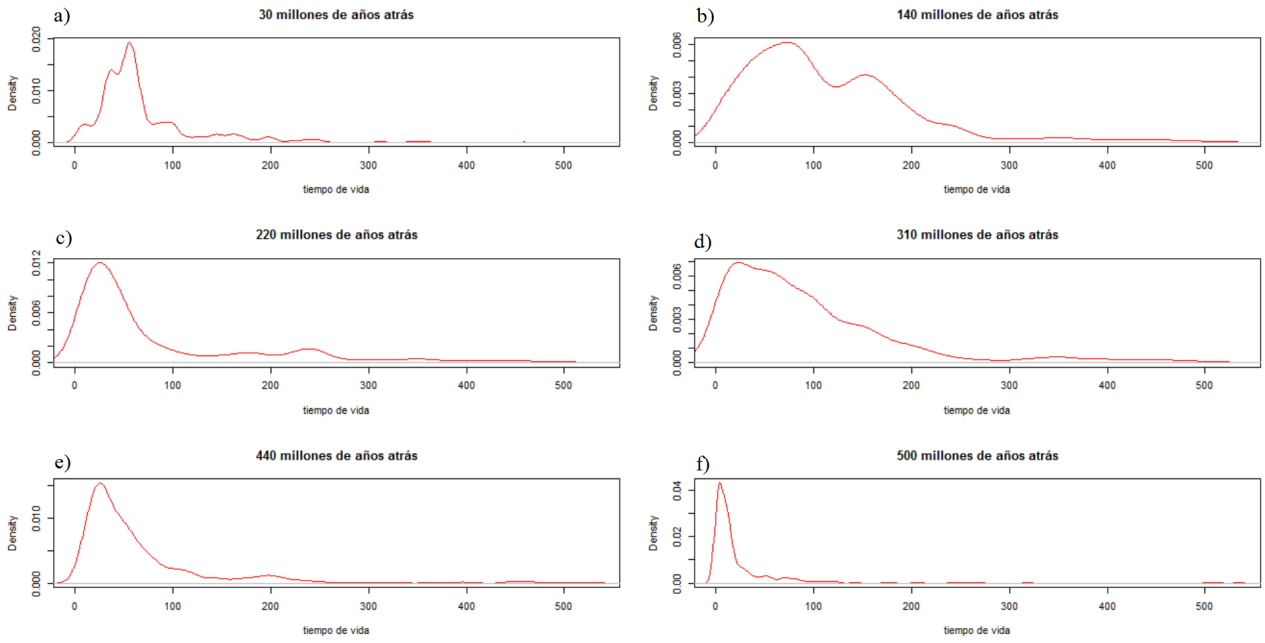

Figura 2.9: Multimodalidad de la densidad de los tiempos de vida para distintos tiempos geológicos. El tablero se puede consultar en https://orozcopedro.shinyapps.io/TiemposKernel/

vida que se expresan en el registro fósil son el resultado de eventos condicionales, entonces la distribución de los tiempos de vida la podemos fijar en exponencial y debido a los procesos de confusión que afectan al registro fósil entonces la exponencialidad se puede distorsionar. La idea de la distribución del registro fósil como una distribución condicional se abordará con mayor profundidad en el siguiente capítulo y quedará clara la razón por la cual la multimodalidad no es necesariamente un impedimento para que los tiempos de vida sean exponenciales.

Problemas en los tiempos geológicos antiguos y recientes

Una característica que se observa en las Figuras 2.8 y 2.9 es que hay mucha diferencia entre los casos (a) y (f) a comparación de todos los demás. Esto implica que cuando nos fijamos en los tiempos geológicos más antiguos y los más recientes entonces los tiempos de vida se comportan muy diferente al resto de los tiempos geológicos. En el caso de los tiempos geológicos más recientes la explicación más probable es el hecho que estos tiempos tienen mezclados los dos tipos de censura que se mencionaron anteriormente. Por otro lado, la diferencia de comportamiento que se observa en los tiempos geológicos antiguos tiene que ver con que no se tiene registro de especies que hayan tenido longevidad mediana. A lo que nos referimos con longevidad mediana es que se no se tienen fósiles de especies que hayan aparecido en tiempos geológicos muy antiguos pero que hayan existido más de 100 millones de años y menos de 300 millones de años.

Notemos en la Figura 2.9 que para todos los casos se puede apreciar un valle en la densidad justo donde están las longevidades medianas. Esto nos puede estar diciendo que la probabilidad de encontrar fósiles de especies que

existieron durante un periodo menor a 300 millones de años y mayor a 100 millones de años es por alguna razón menor que la de encontrar fósiles más longevos o que hay un umbral de supervivencia. Es decir, que cuando una especie tiene entre 100 millones y 300 millones de años de existir entonces su probabilidad de extinción aumenta. Esto no es congruente con el resto del análisis exploratorio ni con lo que se piensa en la teoría de evolución por lo que es más razonable considerar que hay algún factor que provoca que se tenga más información de especies longevas que de especies de longevidad mediana.

2.2.2. Laboratorio de exploración de modelos

El laboratorio de exploración de modelos surgió a partir de la gran utilidad que tuvo la herramienta interactiva de visualización de datos Shiny. Este laboratorio esta montado en Shiny y es un esqueleto de código fácil de modificar que tiene como objetivo proponer un modelo estocástico que represente la evolución y otro que represente los procesos de confusión del registro fósil y a partir de estos simular escenarios acerca de cómo se refleja la extinción en el registro fósil para ser validados con la base de datos que proporcionó Pablo del Monte. En el siguiente capítulo se profundizará acerca de los detalles del laboratorio de exploración de modelos.

2.3. Conclusiones del análisis exploratorio

El objetivo de cualquier análisis exploratorio es entender mejor el problema que se quiere modelar a partir de los datos. A continuación se va a hacer un resumen de las características de los datos que se van a tomar en cuenta en la modelación.

1. El tiempo geológico es muy importante y debe ser tomado en cuenta en cualquier modelo sin importar la sencillez del mismo por lo cual los parámetros del modelo son funciones que dependen del tiempo geológico.

2. Los procesos de confusión son de dos tipos: fenómenos naturales que cambiaron la evolución o la fosilización y la participación del hombre en la recolección de los datos. De cualquier forma son factores que no pueden ser ignorados en el modelo.

3. Modelar la multimodalidad de los tiempos de vida como una probabilidad de encuentro es más razonable que pensarla como reflejo de un fenómeno evolutivo. En otras palabras, los umbrales de supervivencia no los sustenta el análisis exploratorio ni las colaboraciones con Pablo del Monte.

4. Se puede usar el registro fósil como información para validar simulaciones que involucren la interacción entre la evolución y los procesos de confusión.

Capítulo 3

Modelación

Este capítulo consiste en proponer un modelo probabilístico que refleje de manera sensata la información que aporta el registro fósil acerca de la extinción. Los elementos a considerar son: la multimodalidad, los factores de confusión, la exponencialidad genérica y la dependencia del tiempo. Sin embargo, primero haremos una descripción de la concepción general que tenemos de los procesos que afectan la información de la extinción que se refleja en el registro fósil. A partir de esta descripción enunciaremos los ingredientes teóricos que utilizaremos para cuantificar la incertidumbre de estos procesos. Por último, elaboraremos simulaciones que muestran que la concepción general es razonable y útil para dimensionar la complejidad estadística del problema.

3.1. Condiciones propias del contexto

Para comprender mejor la relación que existe entre el registro fósil y la extinción recurriremos a la Figura 3.1. La figura muestra cuadros secuenciales que ilustran los componentes que participan en el proceso de formación del registro fósil a partir de la evolución. Estos componentes se detallan a continuación.

Cuadro I. Pensemos que la evolución de todas las especies es representada mediante estructuras arboladas. Ahora bien, en este primer cuadro se ilustra un escenario a manera de ejemplo en el cual hay tres especies madres (señaladas con un punto verde en la Figura 3.1) que surgieron en distintos momentos de la historia. Cada una de ellas ha sido distinta en términos de cantidad de descendientes y tiempo medio de existencia de sus descendientes en el planeta. Si el registro fósil reflejara toda la información de la extinción entonces podríamos intentar reconstruir exactamente este cuadro con la base de datos que tenemos.

Cuadro II. Este cuadro ilustra que las condiciones climáticas del planeta afectan el proceso de fosilización de las especies. Se muestran con franjas verdes los momentos en los que las condiciones climáticas eran ideales para que la fosilización se llevara a cabo. Por otro lado, con franjas rojas son momentos en los que las condiciones climáticas fueron adversas para la fosilización. Donde no hay franjas suponemos condiciones climáticas intermedias.

Cuadro III. No sólo las condiciones climáticas afectan la fosilización, sino también la estructura física de las especies. Esto

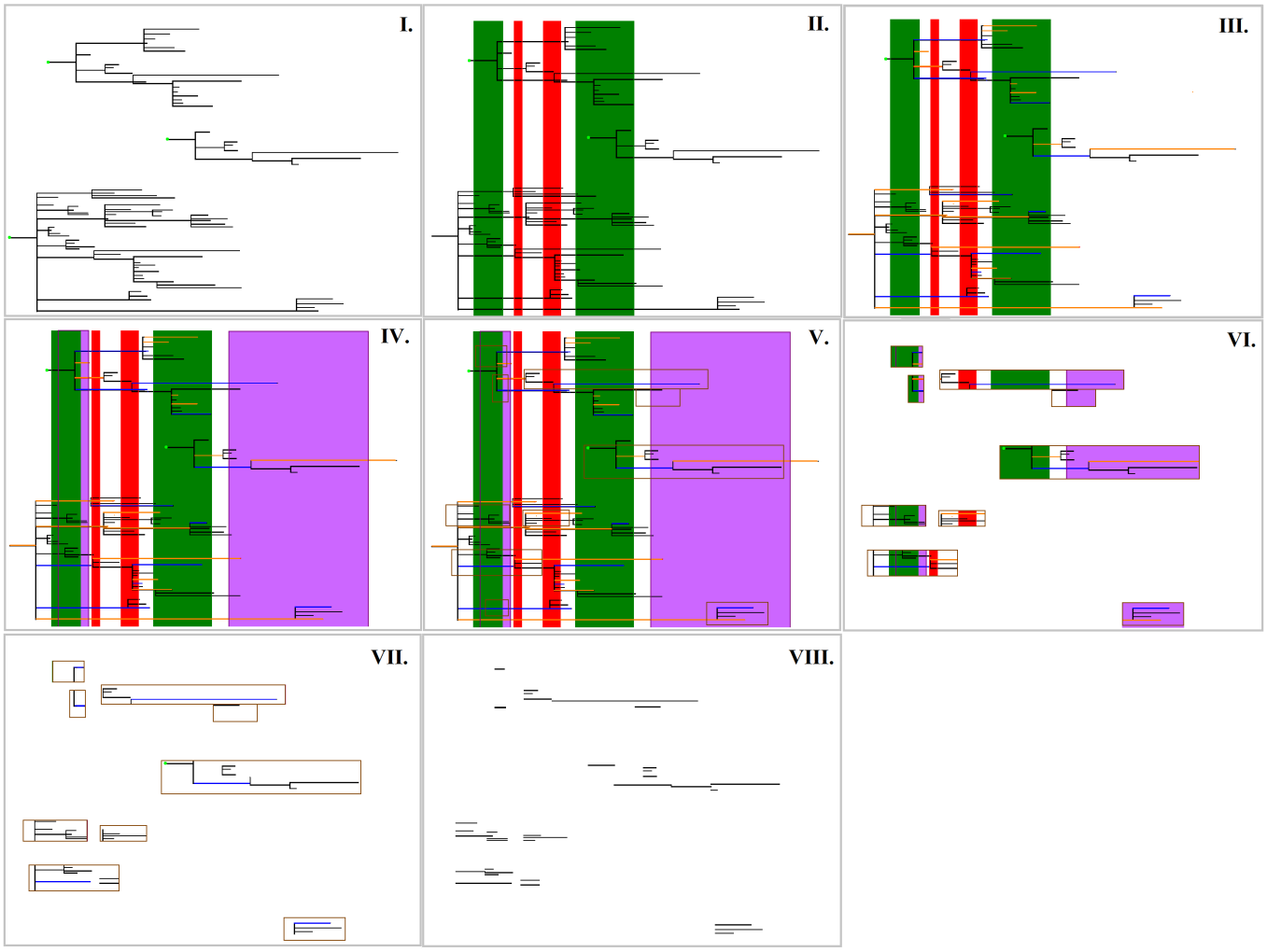

Figura 3.1: Comic que presenta en ocho pasos la relación que existe entre la extinción y el registro fósil que recolecta el hombre.

quiere decir que las especies cuya estructura física es blanda, como por ejemplo los moluscos, nunca van a dejar huella en el registro fósil. Con color naranja se muestran las especies recién mencionadas que tienen estructuras blandas y no van a aparecer nunca en el registro fósil, con color azul las especies que tienen estructuras ideales para la fosilización y con color negro las especies con una estructura intermedia.

Hasta este momento aquellas especies que sean segmentos de color azul y que intercepten en franjas verdes fosilizarán con toda seguridad. Sin embargo, no es suficiente que fosilicen para que se encuentren en la base de datos; hace falta que el fósil sea encontrado.

Cuadro IV. En este cuadro se muestran dos franjas de color morado y lo que quieren mostrar son aquellos tiempos geológicos en los que el ser humano ha realizado mayor cantidad de excavaciones. Esto puede ser por tratarse de capas geológicas que se encuentran más próximas a la superficie, capas geológicas en las que existieron especies que de alguna manera son populares (como los dinosaurios) y generan un mayor interés, etcétera.

Cuadro V. No basta con excavar una capa geológica para asegurar que se van a encontrar todos sus fósiles. Por eso, en este cuadro se muestran rectángulos que encierran algunos segmentos. Lo que representan estos rectángulos son los fósiles que se encontraron y por lo tanto ya forman parte de la base de datos del registro fósil.

Hasta el momento la imagen es muy confusa ya que tiene muchos conceptos sobrepuestos. Por ejemplo, observemos el árbol evolutivo de la especie madre que se encuentra en la parte inferior del cuadro. Ahora concentremos nuestra atención al cuadro demayor tamaño, el cual se encuentra a la altura media del árbol evolutivo. En este cuadro podemos notar todos los elementos hasta el momento mencionados y por lo tanto su interpretación es muy extensa. De hecho, se necesitaron cuatro cuadros anteriores para poder entender toda la información que plasma el cuadro. Tomemos en cuenta que estamos hablando unicamente de una fracción de la figura que a su vez sólo contiene tres especies madres y además la cantidad de factores de confusión que se ilustran son muy pocos a comparación de los que se podrían enunciar en un análisis más profundo. Sin duda alguna es una simplificación de la realidad y es muy probable que haya más procesos que afectan el proceso de fosilización, por lo que en la imagen no es tan confusa como debería ser. Es aún más importante notar que la imagen está repleta de información que el registro fósil no posee.

Cuadro VI. Aquí mostramos lo que el registro fósil sería capaz de representar en un escenario ideal. Con esto nos referimos a que todo lo que está fuera de los cuadros no se ha encontrado y por lo tanto lo borramos de la imagen para ilustrar que no formaría parte del registro fósil. Sin embargo, se muestran elementos que no se reflejan en el registro fósil con el objetivo de mostrar dicho escenario ideal. Este escenario corresponde al caso en que al momento de encontrar fósiles e incorporarlos al registro fósil se estuviera también encontrando información de todos los elementos que se han mencionado en los cuadros anteriores.

Cuadro VII. Sin embargo el registro fósil no tiene información de las condiciones climáticas ni del esfuerzo de muestreo, ni de las estructuras de las especies. Es por eso que este cuadro contiene mucha menos información que el anterior aunque ilustra de manera realista la cantidad de información que el registro fósil posee acerca del proceso de evolución. Como observación adicional notemos que en este cuadro mostramos sólo las especies de las que sí se encontraron fósiles y por eso se borraron los segmentos naranjas que el cuadro VI aún mantenía.

Cuadro VIII. En el cuadro anterior se muestran las especies que se encuentran en el registro fósil y algunas conexiones de descendencia. Encontrar los descendientes es una tarea realizable pero sumamente laboriosa y por lo tanto de la mayoría de los fósiles no se conocen sus descendientes o ancestros. Por eso en este último cuadro se borran las conexiones que se tienen entre las líneas y se muestran todas independientes. Este cuadro es una ilustración fidedigna de la información que posee el registro fósil en relación a la evolución.

Si comparamos los cuadros I y VIII podríamos concluir que salvo por que se perdieron las estructuras arboladas no existe una diferencia tan grande entre la evolución y lo que se refleja en el registro fósil. Pero si ahora observamos el cuadro V y VIII es cuando notamos que hay una gran cantidad de información que esta afectando al registro fósil que no podemos ignorar. Todos estos procesos de confusión los vamos a tratar de capturar con componentes probabilísticos para tomar en cuenta la incertidumbre que aportan al registro fósil. Para poder hacer esto era crucial describir de forma conceptual cuál es la historia que queremos contar con el modelo probabilístico. En resumen, las condiciones propias del contexto son: la evolución como estructura múltiple arbolada, la fosilización y el muestreo incompleto.

3.2. Ingredientes Técnicos

Los ingredientes técnicos son las herramientas estadísticas que vamos a utilizar para modelar la relación entre la evolución y el registro fósil descrito en la Figura 3.1. Dichas herramientas las vamos a clasificar en dos categorías: ingredientes técnicos generales e ingredientes técnicos específicos. Ambos buscan dar una propuesta para la densidad de los tiempos de vida que dependa del tiempo geológico. Esto quiere decir que de manera indirecta van a especificar una función de riesgo de extinción para cada tiempo geológico. Es importante mencionar que los ingredientes generales van a englobar a los ingredientes específicos, es decir que los ingredientes específicos son casos particulares de los ingredientes generales.

3.2.1. Ingredientes Generales

Estos ingredientes son aquellos que determinan el conjunto de herramientas estadísticas o matemáticas que se van a utilizar. Los llamamos generales ya que para poder ser utilizados hace falta especificar muchos de sus elementos. Sin embargo, son suficientes para comprender de forma conceptual el enfoque que se está tomando en la modelación. Es por esto que modelos con especificaciones técnicas complejas y otros con especificaciones técnicas simples pueden tener los mismos ingredientes técnicos generales.

Mezcla de Procesos de ramificación

Lo primero que haremos es concebir que el cuadro I de la Figura 3.1 se podría modelar mediante un conjunto de procesos de ramificación. Decimos que se utiliza un conjunto de procesos de ramificación porque supondremos que hay un proceso de ramificación para cada especie madre, por ejemplo para el caso del cuadro I se tendrían tres procesos de ramificación, uno por cada especie madre (puntos verdes en la Figura 3.1. De esta manera, el cuadro VIII nos estaría diciendo que el registro fósil es un conjunto de tiempos de vida que provienen de procesos de ramificación diversos. Esto es decir que los tiempos de vida vienen de distintas distribuciones. Sin embargo no sabemos de cuántas distribuciones

ni podemos agrupar los tiempos de vida por su distribución. Como metáfora pensemos que tenemos un bosque donde cada árbol representa un proceso de ramificación y queremos estimar la distribución con la que las ramas crecen. Sin embargo nosotros sólo tenemos un camión lleno de ramas y no sabemos de cuantos árboles provienen ni tenemos las ramas agrupadas de ninguna manera.

Esta concepción requiere entonces de una manera de modelar la incertidumbre que capture el hecho de que desconocemos de qué proceso de ramificación proviene cada especie. Volviendo a la metáfora del bosque lo que se hará es pensar en que hubo un mecanismo que "rompió" el bosque dejando sólo algunas de las ramas de los árboles. La manera en que se distribuyen los fragmentos del bosque, es decir las ramas de los árboles incompletos que quedaron después de este rompimiento se puede modelar mediante una distribución de probabilidad. Entonces la distribución de las ramas sería una mezcla entre la distribución del rompimiento y la distribución de la longitud de las ramas. Es decir, que nuestro camión de ramas nos da información de dos distribuciones: la distribución de los tiempos de vida y la distribución de los fragmentos del bosque. Dicho de otra manera, el camión de ramas es el resultado de la mezcla de un número no especificado de procesos de ramificación mediante la distribución del rompimiento de los árboles.

La elección de cada una de estas dos distribuciones será la forma de especificar el modelo final. Un ejemplo muy general de cómo especificar estas dos distribuciones sería la siguiente. Sea $f(x;\lambda)$ la densidad de los tiempos de vida y sea $g(\lambda;\tau)$ la densidad de los fragmentos. Es decir, el parámetro λ es el que determina de qué proceso de ramificación provino un tiempo de vida específico. De esta manera proponemos que existe una infinidad de posibles procesos de ramificación y los tiempos de vida de todos provienen de la misma familia de distribución indexada por λ. Es importante notar que el modelo de fragmentación podría ser mucho más complejo o mucho más sencillo que utilizar una densidad.

Hasta ahora no hemos mencionado dónde está la dependencia del tiempo. Éste será capturado por el parámetro de la distribución de los fragmentos, es decir que indirectamente los parámetros de la distribución de tiempos de vida van a depender del tiempo. Para ser específicos, la distribución de los fragmentos se definirá como $g(\lambda;\tau(t))$, donde t representa el tiempo geológico. Así, para un tiempo geológico fijo t la densidad de los tiempos de vida estará dada por:

$$\tilde{f}(s;\tau(t)) = \int f(s;\lambda)g(\lambda;\tau(t))d\lambda, \tag{3.1}$$

donde s representa el tiempo de vida o longevidad, t el tiempo geológico que estamos observando, $f(s;\lambda)$ es la densidad de los tiempos de vida del proceso de ramificación asociado con λ y $g(\lambda;\tau(t))$ es la densidad del proceso de fragmentación, el cual vamos a suponer independiente de los procesos de ramificación. Resumiendo, tenemos una densidad de tiempos de vida que proviene de la fragmentación de procesos de ramificación.

Probabilidad de detección

La mezcla de procesos de ramificación la utilizamos para modelar la Figura 3.1, y captura la información de los cuadros II y III en la distribución de la fragmentación. Sin embargo, no es razonable suponer que también captura la influencia que tiene el ser humano en la recolección de los fósiles. Para incluir este factor de confusión será necesario incluir la probabilidad de detección como un ingrediente técnico. Notemos que este ingrediente se centra en el hecho de que la especie sí fosilizó, por lo que está bien diferenciado con la fragmentación de los procesos de ramificación. Este

ingrediente técnico va a capturar la información que se ilustra en los cuadros IV al VII. Notemos que estos cuadros ilustran el proceso de que un fósil sea encontrado de aquí el nombre del ingrediente.

La probabilidad de detección del fósil de una especie en particular la concebiremos como una función paramétrica que depende del tiempo de vida. Los parámetros de la función van a depender del tiempo geológico y es mediante éstos que podemos elegir de qué manera se incluirán los factores de confusión. Vamos a denotar a la probabilidad de detección de la siguiente manera:

$$p(s;\beta(t)), \tag{3.2}$$

donde t representa el tiempo geológico, s el tiempo de vida de alguna especie y β es el parámetro de la función. La interpretación del parámetro β va a depender de los elementos que queremos se reflejen en la función p. Por ejemplo, si la función p se elije de forma que represente la forma en que el cambio climático afecte al registro fósil entonces el parámetro β podría ser la temperatura promedio de la era geológica por ejemplo. Para el caso de esta tesis la interpretación del parámetro β será más fácil de entender cuando se hable de los componentes técnicos específicos del modelo y se mencione la justificación de la elección de p. Esta expresión sugiere que la longevidad de una especie va a determinar la probabilidad de que el fósil sea encontrado. Más adelante desarrollaremos con mayor detalle esta idea.

Muestreo por encuentro

El problema de que el registro fósil sea resultado de procesos de confusión está ligado en parte a que los datos que lo componen no se obtienen mediante un experimento planeado, es decir que los datos no son observados de una manera sistemática o controlada. En ecología esto es muy común y se denomina muestreo por encuentro [17]. Al obtener datos de esta manera se produce un sesgo, el cual depende del mecanismo de observación. Dicho de forma más natural, los datos no tienen una probabilidad uniforme de ser observados. Esta probabilidad de ser observado es la probabilidad de detección que se mencionó antes.

Una manera de entender el muestreo por encuentro es la siguiente. Supongamos que tenemos la variable aleatoria S con densidad $\tilde{f}(s;\tau)$ que modela el tiempo de vida de una especie. Ahora supongamos que tenemos una variable Bernoulli D que modela la probabilidad de que el fósil de una especie con tiempo de vida s sea encontrado (definido como éxito) y esta variable tiene como probabilidad $p(s;\beta)$, es decir que la probabilidad de éxito depende del tiempo de vida. Bajo esta notación $P(D=1|S=s)=p(s;\beta)$ y $P(S=s)=\tilde{f}(s;\tau)$. Lo que nos interesa entonces conocer es la probabilidad de que un tiempo de vida s sea observado dado que se encuentra en el registro fósil. Por teorema de Bayes tenemos la siguiente relación:

$$P(S=s|D=1)=\frac{p(s;\beta)\tilde{f}(s;\tau)}{\int p(s;\beta)\tilde{f}(s;\tau)ds}. \tag{3.3}$$

Retomando nuestra propuesta para modelar la distribución de los tiempos de vida de las especies con la expresión 3.1, entonces ahora podemos juntar las ideas de mezclar procesos de ramificación con la probabilidad de detección y muestreo por encuentro. El modelo general que involucra los procesos de confusión está determinado por

$$f^*(s;\theta(t))=\frac{p(s;\beta(t))\tilde{f}(s;\tau(t))}{\int p(s;\beta(t))\tilde{f}(s;\tau(t))ds}, \tag{3.4}$$

donde $\theta(t)$ son los parámetros de todo el modelo, es decir $\theta(t) = (\beta(t), \tau(t))$. Es decir, que f^* es la densidad de los tiempos de vida de las especies que se encuentran en el registro fósil tomando en cuenta los factores de confusión.

3.2.2. Ingredientes Técnicos Específicos

Los ingredientes técnicos específicos surgirán al especificar los elementos de la expresión 3.4. Se investigarán tres opciones diferentes de 3.1. La primera de ellas va a estar justificada por la génesis de la distribución mezcladora. La segunda se propuso para disminuir problemas numéricos que surgen al calcular integrales iteradas. Por último la tercera propuesta es una generalización de la segunda para poder darle mayor flexibilidad al modelo. En el caso de la segunda componente del modelo, es decir la expresión 3.2, presentaremos una única propuesta basada en pláticas con el Dr. Del Monte. Esta propuesta tiene como objetivo reflejar la intuición que se tiene acerca de la frecuencia con la que las especies son encontradas en el registro fósil con base en su longevidad. Para poder verificar si la elección de estos ingredientes es adecuada se van a hacer ejercicios de simulación, los cuales mencionaremos más adelante.

Elección de distribución de tiempos de vida de los procesos de ramificación

La distribución de los tiempos de vida es el elemento de nuestro modelo del cual tenemos mayor información gracias al registro fósil. Una de las conclusiones del análisis exploratorio es que existe una aparente exponencialidad en los tiempos de vida. Supondremos que la distribución de los tiempos de vida es exponencial. Sin embargo los procesos de confusión provocan que el análisis exploratorio no dé suficiente evidencia para determinar que los tiempos de vida se distribuyan de manera exponencial. Por esta razón propondremos que los tiempos de vida del conjunto de procesos de ramificación tengan distribución exponencial. Es decir, que cada proceso de ramificación es especificado mediante la elección del parámetro de la distribución de sus tiempos de vida λ. De esta forma estamos permitiendo que exista una infinidad de procesos de ramificación ya que el espacio parametral es $\mathbb{R}^+$.

Para determinar a un proceso de ramificación es necesario especificar dos distribuciones. En nuestro caso no tenemos información acerca de los descendientes y ancestros de las especies que se reflejan en el registro fósil. Es por esta razón que no podemos proponer una distribución de los nacimientos para los procesos de ramificación. Sin embargo, recordemos que no es necesario para nuestro estudio especificar dicha distribución ya que el objetivo es hacer inferencia en la tasa de extinción y para eso basta conocer la distribución de los tiempos de vida.

Distribución del proceso de rompimiento

Para completar la expresión 3.1 hace falta especificar la distribución que rige el proceso de fragmentación de los procesos de ramificación. Es en esta componente del modelo donde daremos tres opciones distintas debido a que aquella cuya justificación es más susceptible a controversia.

Propuesta I.

Pensaremos en una familia de distribuciones que tenga una génesis que represente el rompimiento de los procesos de ramificación. Àitchison y Brown [18] mencionan que la distribución Lognormal de dos parámetros refleja la distribución de los pedazos resultantes de romper natural o artificialmente un mineral. Supondremos que la fosilización es el

resultado de un rompimiento de la evolución y por lo tanto la elección de $g(\lambda;\tau(t))$ como la densidad de una variable aleatoria Lognormal se basa en la génesis que Àitchison y Brown describen.

De esta manera tenemos que la primera propuesta para la distribución de los tiempos de vida es una mezcla infinita de distribuciones exponenciales donde los pesos se distribuyen mediante una distribución Lognormal. Por ejemplo, supongamos que conocemos los parámetros $\tau(t)$ de la distribución Lognormal para un tiempo fijo t. Entonces esto nos define una variable aleatoria Λ con densidad $g(\lambda;\tau(t))$. Un proceso de ramificación se especifica a raíz de que Λ tome un valor, es decir que si X es un tiempo de vida entonces su distribución es condicional a que pertenezca a un proceso de ramificación indexado por $\Lambda = \lambda$. Por lo tanto la expresión 3.1 queda de la siguiente manera:

$$\tilde{f}(s;\tau(t)) = \int_{\mathbb{R}^+} \frac{1}{\lambda^2\sigma_t\sqrt{2\pi}} \exp\left\{\frac{-(\ln(\lambda)-\mu_t)^2}{2\sigma_t^2} - \frac{s}{\lambda}\right\} d\lambda, \tag{3.5}$$

donde $\tau(t) = (\mu_t, \sigma_t)$ son los parámetros de la Lognormal. En la sección de simulación especificaremos como estimaremos a $\tau(t)$.

Propuesta II.

La propuesta 3.5 es una integral que no se puede resolver de manera analítica. De esto que la segunda propuesta se centre en la necesidad práctica de contar con una expresión analítica cerrada para la densidad de tiempos de vida. Es por eso que se propuso que la distribución del rompimiento de los procesos de ramificación sea exponencial. De esta manera, $g(\lambda;\tau(t))$ en este caso es la densidad de una distribución exponencial. Es decir tenemos que

$$\tilde{f}(s;\tau(t)) = \frac{\gamma_t}{(\gamma_t + s)^2}, \tag{3.6}$$

donde $\tau(t) = \gamma_t$ es el parámetro de intensidad de la distribución de Λ. En esta propuesta ya tenemos una expresión analítica cerrada de modo que en el cálculo de 3.4 ya sólo se debe realizar una integral de manera numérica en lugar de una integral anidada como con la Propuesta I.

Propuesta III.

La propuesta II se puede hacer más flexible si en lugar de utilizar una distribución exponencial para el rompimiento usamos una distribución Gama. Esto es decir que ahora vamos a declarar que $\Lambda \sim \text{Gama}(\alpha, \beta)$. De esta manera $g(\lambda, \tau(t))$ es la densidad de una variable aleatoria Gama. La expresión queda de la siguiente manera:

$$\tilde{f}(s;\tau(t)) = \frac{\alpha_t^{\theta_t}\theta_t}{(\alpha_t + s)^{\theta_t+1}}, \tag{3.7}$$

donde $\tau(t) = (\alpha_t, \theta_t)$ son los parámetros de intensidad y forma, respectivamente, de la distribución Gama.

Cabe especificar que la parametrización de la exponencial fueron distintas en el caso de la propuesta I y las propuestas II y III. Para el caso de la propuesta I la parametrización se centró en hacer más estable numéricamente el cálculo de 3.5. Para las propuestas II y III se tomó una parametrización que simplificó el cálculo analítico de 3.6 y 3.7.

Probabilidad de detección

La probabilidad de detección es un componente que puede poseer una interpretación muy rica en términos de otros problemas ajenos a la extinción. Por ejemplo, si se tiene información acerca del cambio climático y se puede incluir en la expresión 3.2 entonces los parámetros tendrán interpretación del cambio climático. Para nuestro caso, las propuestas para 3.2 estarán basadas principalmente en la intuición ganada en las pláticas con Pablo del Monte y lo aprendido en el análisis exploratorio. Lo más importante que se va a querer reflejar con nuestra propuesta de probabilidad detección es el hecho de que los fósiles de especies longevas tienden a ser los más encontrados. Esto lo suponemos por el hecho de que mayor tiempo de permanencia en el planeta asegura que habrá más oportunidades de fosilizar a comparación de las especies que permanecen poco tiempo en el planeta. Por otro lado las especies no longevas son muy comunes por lo que esto afecta también la probabilidad de que sean vistas en el registro fósil. Esta dualidad propone que la expresión 3.2 tenga un comportamiento que refleje una probabilidad baja para especies de tiempo de vida corto y que esta probabilidad vaya en aumento con tasa creciente hasta un cierto punto (no determinado) en el cual la probabilidad siga creciendo pero con una tasa decreciente (ver Figura 3.2). De esta forma vamos a incluir dos parámetros en la probabilidad de detección. La función tiene la siguiente expresión:

$$p(s; a_t, b_t) = \frac{e^{a_t(s-b_t)}}{1 + e^{a_t(s-b_t)}} c_1 + c_2, \tag{3.8}$$

donde $c_1 + c_2$ es igual al límite superior de la probabilidad de detección, c_2 es el límite inferior de la probabilidad de detección, a_t es una constante que está relacionada con la tasa a la que aumenta la probabilidad de detección y b_t es un parámetro de centralidad que determina el punto en el cual la tasa a la que aumenta la probabilidad de detección comienza a ser decreciente. Esta propuesta de probabilidad de detección tiene un número excesivo de parámetros y provoca que la verosimilitud sea plana por lo que las constantes c_1 y c_2 se fijaron en 0.9 y 0.05 respectivamente. Es por esto que sólo se denota a $\beta_t = (a_t, b_t)$ como un parámetro de la expresión 3.8.

El rol que juega β es muy importante en el sentido que puede mejorar la forma de modelar los factores de confusión conforme se tenga mayor información de los mismos. En nuestro caso no se tiene información de los factores de confusión. Aunque no se aprecia en la expresión 3.2, la probabilidad de detección puede ser un modelo complejo por sí mismo.

Propuesta

Juntando las expresiones 3.5, 3.6, 3.7 y 3.8 para sustituir en 3.4 tenemos las siguientes tres propuestas.

I. Distribución del rompimiento Lognormal

$$f^*(s; \theta(t)) = \frac{\left[\frac{e^{a_t(s-b_t)}}{1 + e^{a_t(s-b_t)}} c_1 + c_2\right] \left[\int_{\mathbb{R}^+} \frac{1}{\lambda^2 \sigma_t \sqrt{2\pi}} \exp\left\{\frac{-(\ln(\lambda) - \mu_t)^2}{2\sigma_t^2} - \frac{s}{\lambda}\right\} d\lambda\right]}{\int \left[\frac{e^{a_t(s-b_t)}}{1 + e^{a_t(s-b_t)}} c_1 + c_2\right] \left[\int_{\mathbb{R}^+} \frac{1}{\lambda^2 \sigma_t \sqrt{2\pi}} \exp\left\{\frac{-(\ln(\lambda) - \mu_t)^2}{2\sigma_t^2} - \frac{s}{\lambda}\right\} d\lambda\right] ds}, \tag{3.9}$$

donde $\theta(t) = (\mu_t, \sigma_t, b_t, a_t)$.

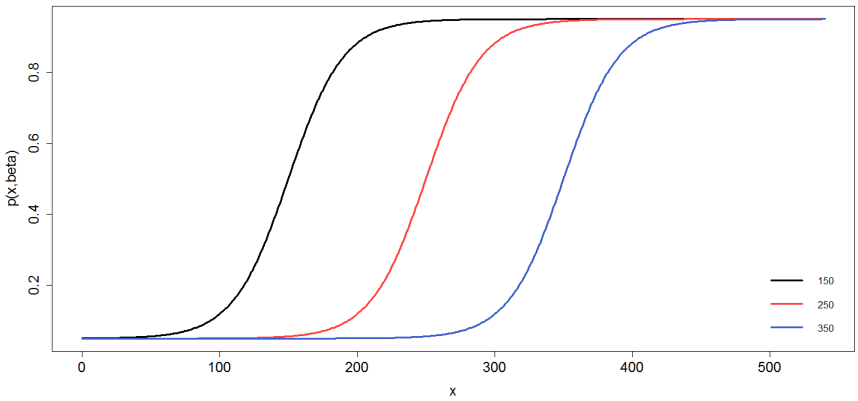

Figura 3.2: Tres ejemplos de probabilidad de detección. Estas tres opciones dependen de un único parámetro de localización con los siguientes valores: 150 en negro, 250 en rojo y 350 en azul. En esta imagen, $k = 0.05$, $c_1 = 0.9$ y $c_2 = 0.05$

II. Distribución del rompimiento Exponencial

$$f^*(s;\theta(t)) = \frac{\left[\frac{e^{a_t(s-b_t)}}{1+e^{a_t(s-b_t)}}c_1 + c_2\right]\left[\frac{\gamma_t}{(\gamma_t+s)^2}\right]}{\int\limits_{\mathbb{R}^+}\left[\frac{e^{a_t(s-b_t)}}{1+e^{a_t(s-b_t)}}c_1 + c_2\right]\left[\frac{\gamma_t}{(\gamma_t+s)^2}\right]ds}, \tag{3.10}$$

donde $\theta(t) = (\gamma_t, b_t, a_t)$.

III. Distribución del rompimiento Gama

$$f^*(s;\theta(t)) = \frac{\left[\frac{e^{a_t(s-b_t)}}{1+e^{a_t(s-b_t)}}c_1 + c_2\right]\left[\frac{\alpha_t^{\beta_t}\beta_t}{(\alpha_t+s)^{\beta_t+1}}\right]}{\int\limits_{\mathbb{R}^+}\left[\frac{e^{a_t(s-b_t)}}{1+e^{a_t(s-b_t)}}c_1 + c_2\right]\left[\frac{\alpha_t^{\beta_t}\beta_t}{(\alpha_t+s)^{\beta_t+1}}\right]ds}, \tag{3.11}$$

donde $\theta(t) = (\alpha_t, \beta_t, b_t, a_t)$.

3.3. Experimento de simulación

La forma en que utilizaremos los datos para evaluar el modelo es mediante la simulación. Esto significa que con ayuda de una computadora vamos a crear un registro fósil artificial a partir de nuestro modelo y comparar con el registro fósil que tenemos. Los datos que tenemos servirán para dos propósitos: estimar los parámetros del modelo y comparar nuestro registro fósil artificial con el original. En la medida de que sean indistinguibles ambos registros se concluirá que el modelo es razonable.

Para comparar un registro fósil simulado con el real utilizaremos dos métodos gráficos. El primero es obtener los tiempos de vida simulados de nuestro modelo y hacer una gráfica Q-Q contra los tiempos de vida del registro fósil real. El segundo es comparar la gráfica de la densidad estimada por el método kernel de los tiempos de vida del registro fósil real con la gráfica de la densidad teórica del registro fósil simulado. Hay que notar que para esta segunda herramienta no es necesaria la simulación aunque sí se va a usar la computadora para resolver integrales de forma numérica. Para ambos casos se utilizan los datos reales para estimar los parámetros del modelo y poder graficar. Vamos a considerar razonable un modelo que sea satisfactorio a la luz de ambos métodos gráficos.

Para realizar los experimentos se construyó una aplicación de *Shiny* (https://orozcopedro.shinyapps.io/Simulador/) en la cual se pueden controlar los valores de los parámetros para un tiempo geológico fijo. La aplicación también cuenta con un botón para realizar la estimación de los parámetros por máxima verosimilitud, la cual usa parámetros encontrados manualmente como punto de inicio para el algoritmo de optimización. De esta manera se puede elegir un punto de inicio que parezca estar razonablemente cerca del estimador de máxima verosimilitud según las dos gráficas antes mencionadas. El objetivo de permitir la búsqueda de los parámetros de forma manual es tener una manera de evaluar las dificultades para optimizar la función de verosimilitud. Esto quiere decir que se pueden buscar manualmente distintos puntos de inicio razonables desde los cuales se espera obtener el mismo estimador de máxima verosimilitud. En caso de que esto no suceda entonces podemos sospechar que la función de verosimilitud es difícil de optimizar.

Presentaremos un cuadro de resultados para cada experimento. En estos cuadros podremos ver siete tiempos geológicos distintos y para cada tiempo geológico dos simulaciones. En cada simulación se propone un punto de inicio razonable escogido manualmente que debe cumplir que al menos uno de los dos métodos de evaluación gráfica sea razonable. También se presenta el punto al que el método numérico Nelder-Mead [12] converge en la optimización de la función de Logverosimilitud y una valuación subjetiva de los métodos gráficos. Un resultado adecuado sería que para los dos puntos de inicio se converge al mismo punto final, el cual es evaluado positivamente en las dos gráficas. Estos cuadros pretenden ilustrar los problemas y virtudes de cada uno de los experimentos. Debido a que el experimento se evaluó con una herramienta interactiva que no puede ser vista en un documento se recurrió a reportar este resumen de lo encontrado.

Por último, mostraremos tres imágenes de la aplicación *Shiny* que se utilizó para la evaluación de los experimentos. En ellas vamos a mostrar un ejemplo para cada experimento (en el mismo tiempo geológico) en el cual se encontraron de manera manual parámetros que cumplen de manera razonable los dos métodos de evaluación gráfica. Es importante notar que en la parte inferior de estas imágenes se encuentra la función de probabilidad de detección, es decir la ecuación 3.8. Observemos que esta ecuación nos puede dar un comportamiento como el de la Figura 3.4 o como el de las Figuras 3.5 y 3.6, dependiendo del parámetro de localización. Esto último es señal de que el modelo tiene intrínsecamente al menos dos opciones generales de probabilidad de detección.

3.3.1. Experimento I

Para el primer experimento se abordó la expresión 3.9 como modelo a simular. Ninguna de las dos integrales que participan en el modelo tienen solución analítica por lo que vamos a utilizar métodos numéricos para poder solucionarlas. Esta es la principal dificultad que caracteriza a este modelo dado que una de las integrales se encuentra iterada y por lo tanto se debe hacer un gran número de veces. Recordemos que este modelo posee la propiedad de tener

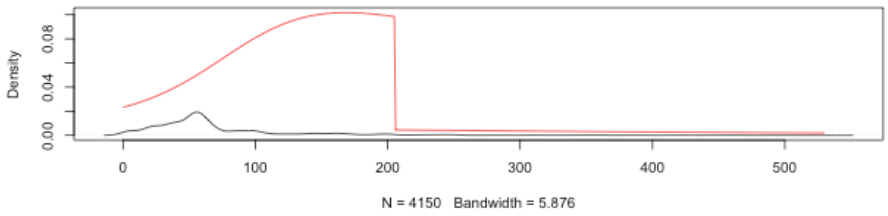

Figura 3.3: Ejemplo de un caso en el que los problemas numéricos se manifiestan al momento de graficar la densidad teórica.

una densidad mezcladora con una génesis muy adecuada y es por esta razón que se evaluó a pesar de las dificultades numéricas que presenta.

En la Figura 3.4 podemos observar un ejemplo de la aplicación de *Shiny* https://orozcopedro.shinyapps.io/Simulador en la cual manualmente se encontró una combinación de parámetros razonable. La densidad teórica (en color rojo) parece estar trasladada hacia la derecha a comparación de la densidad estimada de los datos (en color negro). Además, la densidad estimada de los datos presenta un valle aproximadamente en el tiempo de vida de 100 millones de años y la densidad teórica del modelo no tiene ningún valle. A pesar de esto en la imagen se puede observar que la gráfica QQ tiene muy buen desempeño. Por lo anterior, en efecto se puede suponer que la elección de estos parámetros es razonable.

Como se previó, la optimización de la función de verosimilitud es muy complicada pues se tiene evidencia de que es considerablemente plana (ver Cuadro 3.1), lo cual genera estimación sobre los parámetros de interés provistos de mucha incertidumbre. Es decir, que se necesita un modelo más fino y la incorporación de otras fuentes de información. Por ejemplo datos del cambio climático, de la fosilización, el esfuerzo de muestreo heterogéneo a lo largo de las capas geológicas o de cualquier otro fenómeno que interviene en la inferencia de la tasa de extinción a partir del registro fósil.

En el Cuadro 3.1 podemos observar los resultados para distintos tiempos geológicos. En él se pueden ver los problemas numéricos y de verosimilitud plana. Para verificar que la verosimilitud es plana podemos notar que en ninguno de los casos se convergió al mismo punto final, es decir, que el punto óptimo encontrado por el método numérico depende del punto de arranque. Los problemas numéricos más graves se manifiestan a partir del tiempo geológico 340, en los cuales la Logverosimilitud presentan un valor positivo. Esto no tiene sentido pues la logverosimilitud siempre debe ser negativa (además del caso en el que el método no convergió). Por último, cabe mencionar que en algunos casos la gráfica de la densidad teórica presentaba discontinuidades reflejo de estos problemas numéricos (ver Figura 3.3). Debido a que no se cuenta con acceso a otras fuentes de información se decidió modificar el modelo.

3.3.2. Experimento II

Para el caso de este experimento se tomó la expresión 3.10 como densidad final de los tiempos de vida. Esta propuesta se eligió para poder realizar una de las dos integrales de manera analítica y es considerablemente más sencilla que la propuesta 3.9. De la misma manera que en el Experimento I se utilizó la aplicación *Shiny*

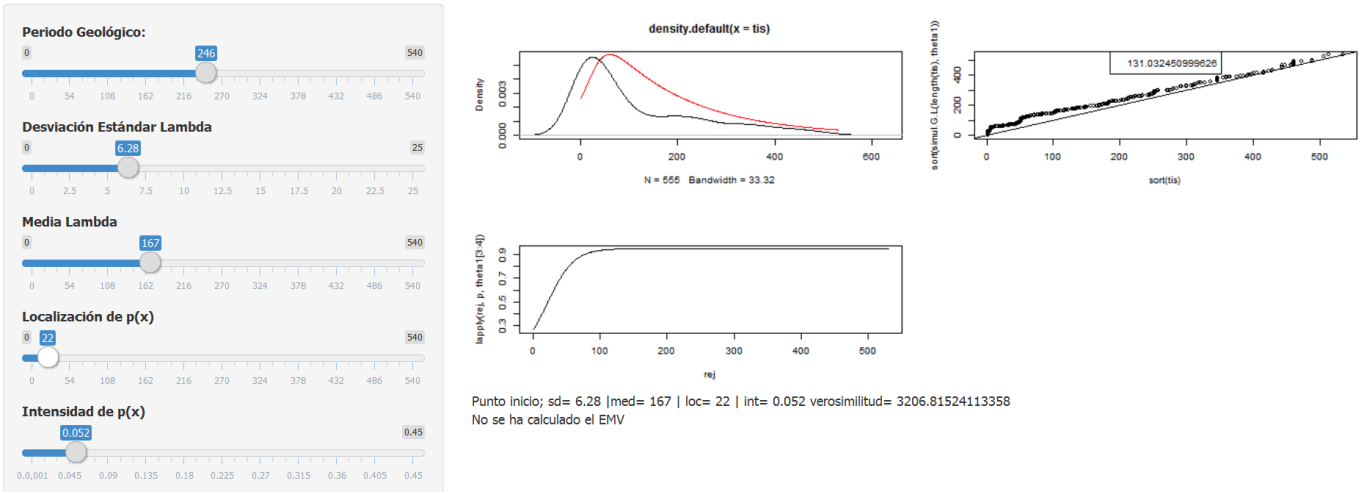

Figura 3.4: Aplicación *Shiny* para el Experimento I sin estimación por máxima verosimilitud. Los tiempos de vida corresponden al tiempo geológico de hace 246 millones de años.

Cuadro 3.1: Resultados Modelo Log Normal

"QQ=si"significa que el qqplot se aproxima razonablemente a una recta de 45 grados, "Dens=si"significa que la densidad teórica se parece a la densidad estimada

T	Sd	Mean	Loc	Int	QQ	Dens	LogVero	Sd	Mean	Loc	Int	QQ	Dens	LogVero
	Punto de inicio							Punto Final						
40	6.28	167	22	0.052	no	si	-21106	2.22	172	22.3	0.001	no	no	-827
	15.45	21	45	0.179	si	no	-22667	11.83	29.77	45.93	0.228	si	si	-20187
115	23.47	37	128	0.08	si	si	-11481	2.32	102	4.03	0.05	si	no	-217
	15.42	128	24	0.165	no	si	-10750	19.97	168.24	24.24	0.03	no	no	-7008
178	11.45	56	183	0.084	no	si	-6220	4.89	47.35	549	0.001	si	no	-1596
	5.47	123	7	0.084	si	si	-6084	11.18	95.63	6.28	0.1	si	no	-4057
279	2.51	89	11	0.084	si	no	-9104	15.21	85.78	12.22	0.16	si	si	-6912
	5.25	22	89	0.084	no	si	-8813	3.23	22.57	144.1	0.21	no	no	-6781
340	11.99	48	10	0.16	si	si	-8339	2.31	102.2	10.29	0.06	si	no	479
	20.5	29	90	0.26	si	no	-9059	El método numérico no convergió						
404	6.89	25	14	0.16	si	si	-9773	1.16	30.4	13.4	0.03	si	no	2816
	1.68	25	11	0.151	si	no	-14046	4.33	39.6	7.15	0.13	si	no	335
450	6.84	25	14	0.16	si	si	-3131	4.1	40.8	13.2	0.05	no	no	250
	8.07	21	19	0.25	si	si	-3441	1.45	84.7	4.7	0.001	no	no	2011

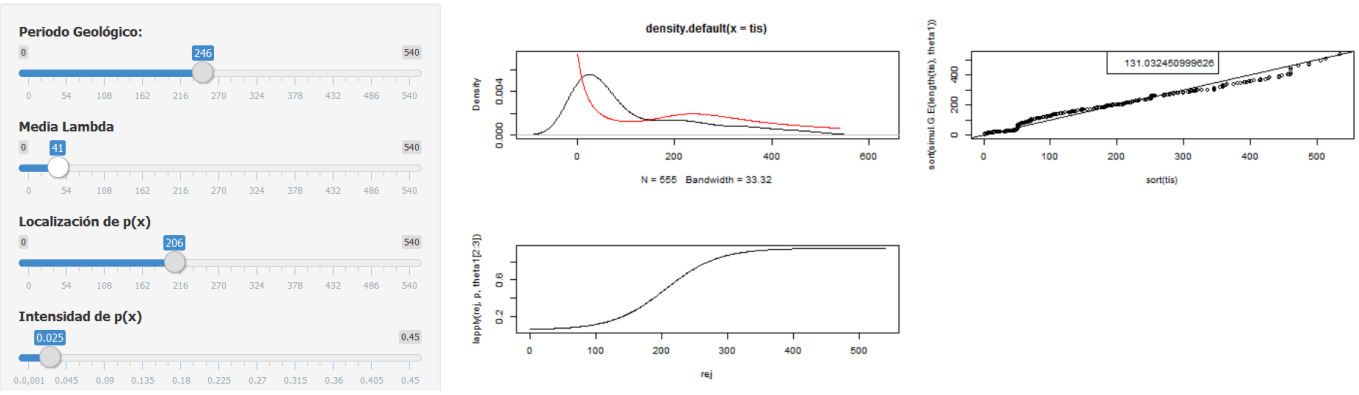

Figura 3.5: Aplicación *Shiny* para el Experimento II sin estimación por máxima verosimilitud. Los tiempos de vida corresponden al tiempo geológico de hace 246 millones de años.

(https://orozcopedro.shinyapps.io/Simulador/) para evaluar la calidad de la propuesta y las posibles dificultades en la estimación de los parámetros.

En la Figura 3.5 se puede ver que la selección manual de los parámetros cumple de forma muy adecuada la gráfica QQ. Para este caso la densidad teórica sí muestra el valle cerca de los 100 millones de años que muestra la densidad estimada con los datos. Sin embargo, la densidad teórica presenta un comportamiento bastante distinto para los tiempos de vida cortos. Experimentando con muchas combinaciones de los parámetros dentro de la aplicación de *Shiny* no se logró encontrar una combinación de parámetros que lograra una densidad teórica adecuada. Nos pareció que para mejorar este problema se debía considerar un modelo que tuviera mayor libertad, y es por esto que una generalización de este mismo modelo podría ser una opción adecuada.

Para este modelo en cinco de los siete tiempos geológicos se puede apreciar que la convergencia el punto final fue la misma sin importar los dos puntos de inicio (ver Cuadro 3.2). Esto nos dice que se logró mejorar la forma de la Logverosimilitud y por lo tanto es más fácil de optimizarla. Sin embargo, sólo en un tiempo geológico el punto final fue razonable según los dos métodos gráficos. Como se mencionó anteriormente, esto posiblemente es reflejo de que el modelo es demasiado restrictivo y aún cuando se encuentra el estimador de máxima verosimilitud mediante optimización numérica, el modelo no es adecuado para los datos en seis de los siete tiempos geológicos. Es por esta razón que reiteramos la necesidad de un modelo diferente, en particular una generalización.

3.3.3. Experimento III

Este experimento corresponde a la expresión 3.11, la cual proviene de considerar una densidad mezcladora Gama. La densidad Gama es una generalización de la densidad Exponencial, y por eso se decidió optar por esta propuesta como un candidato adecuado que respondiera a las dos necesidades creadas por los experimentos I y II: sencillez analítica y mayor libertad en el modelo.

En la Figura 3.6 podemos apreciar que existe al menos un conjunto de parámetros que hace que la densidad teórica sea bastante cercana a la densidad estimada. En el caso de la gráfica QQ consideramos que es bastante buena aunque

Cuadro 3.2: Modelo Exponencial

T	Mean	Loc	Int	QQ	Dens	LogVero	Mean	Loc	Int	QQ	Dens	LogVero
	Punto de inicio						Punto Final					
40	25	17	0.17	si	si	20980	15.7	32.5	0.15	no	no	20760
	41	14	0.303	si	si	21007	15.7	32.5	0.15	no	no	20760
115	94	18	0.032	si	no	10834	25.8	44.4	0.11	si	no	10633
	72	15	0.16	si	si	10759	25.8	44.4	0.11	si	no	10633
178	72	15	0.16	si	si	6143	53.6	14.5	0.13	si	si	6129
	32	147	0.06	si	no	6602	53.6	14.5	0.13	si	si	6129
279	43	28	0.11	si	si	8053	19.2	20.12	0.23	no	si	7896
	98	15	0.35	si	si	8189	19.2	20.12	0.23	no	si	7896
340	45	15	0.35	si	si	8583	26.48	14.12	0.37	no	si	8528
	91	540	0.002	si	si	8977	47.6	540	0.002	si	si	8845
404	21	14	0.27	no	si	9883	14.5	8.7	0.5	no	si	9727
	62	10	0.27	si	si	10230	14.5	8.7	0.5	no	si	9727
450	25	10	0.38	si	si	8029	8.9	10.7	0.5	no	si	7879
	26	453	0.8	si	si	8913	25.7	540	0.5	no	si	8766

definitivamente no ideal o mejor que la de los experimentos I y II. Notemos que la función de probabilidad de detección tiene la forma sigmoidal, al igual que en el caso del Experimento II. Este comportamiento de la función representa mejor la intuición que se tiene a partir de las pláticas con el biólogo Pablo del Monte.

Este experimento presenta problemas numéricos en dos ocasiones, en las cuales el método Nelder-Mead no convergió. Además en ninguno de los siete tiempos geológicos el punto final es el mismo para los dos puntos de inicio, por lo que la Logverosimilitud volvió a ser demasiado plana al igual que en el Experimento I. Es importante mencionar que a pesar de los problemas de forma de la Logverosimilitud el tiempo de cómputo necesario para que el método Nelder-Meade convergiera fue mucho menor que en el caso del Experimento I. Por último debemos mencionar que en cinco ocasiones el método numérico obtuvo una solución frontera en los parámetros de la función de probabilidad de detección. Esto también es señal de los problemas numéricos. Sin embargo, puede ser que la función de probabilidad de detección no es la adecuada ya que las fronteras se definieron para que la función presentara el comportamiento deseado. En dos ocasiones de las cinco, la frontera a la que se convergió corresponde a una función de probabilidad de detección que podría ser remplazada por una función discreta ya que la tasa de crecimiento crece de manera muy acelerada. En los restantes tres casos el punto frontera corresponde a una función de detección constante, es decir que la función no sería útil. Por último nos queda decir que este Experimento no ayudó a resolver ninguno de los problemas presentados por los experimentos anteriores. Sin embargo, es relevante reportarlo pues es una opción natural a partir de los resultados del Experimento II.

Los modelos aquí propuestos no pretenden ser exhaustivos. Podríamos en principio proponer modelos que tengan muchos menos parámetros, por ejemplo una función de detección de un solo parámetro. El motivo por el cuál no

Cuadro 3.3: Modelo Gama

T	Sd	Mean	Loc	Int	QQ	Dens	LogVero	Sd	Mean	Loc	Int	QQ	Dens	LogVero
	Punto de inicio							Punto Final						
40	18	250	29	0.18	si	no	20991	7.35	252.7	37.8	0.12	no	si	20259
	23.7	112	34	0.26	no	si	20548	4.55	123.3	41.1	0.13	no	si	20262
115	4.31	112	129	0.41	si	no	11154	El método numérico no convergió						
	5.13	213	8	0.18	si	si	10683	49.67	3667.17	46	0.05	si	si	10444
178	5.13	213	8	0.18	si	si	6104	5.48	484	0.18	0.02	si	no	6084
	5.68	100	166	0.13	si	no	6429	El método numérico no convergió						
279	22.28	415	13	0.307	si	si	7982	2.77	119.1	15.3	0.21	no	si	7863
	4.81	133	535	0.023	si	no	8451	2.41	131.6	537	0.001	si	si	8018
340	4.81	133	535	0.023	si	si	9147	17.8	1180	1.7	0.007	si	si	8601
	1.1	58	89	0.26	si	no	10288	3.98	193.1	10.37	0.4	no	si	8467
404	25	381	1	0.295	si	si	10039	8.9	379.8	5.3	0.5	no	si	9756
	0.69	40	464	0.02	si	si	11488	1.46	41.12	486.8	0.001	no	si	9954
450	3.05	82	464	0.02	si	si	8854	2.87	85.9	507.6	0.001	no	si	8079
	25	386	6	0.314	si	si	8386	11.63	386.2	7	0.5	no	si	7924

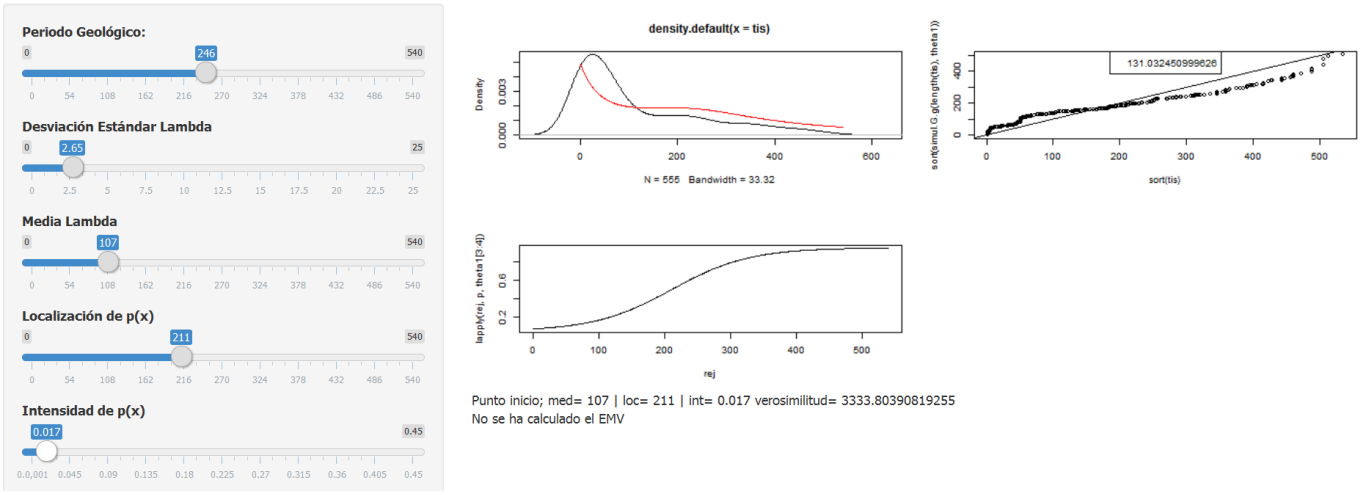

Figura 3.6: Aplicación *Shiny* para el Experimento III sin estimación por máxima verosimilitud. Los tiempos de vida corresponden al tiempo geológico de hace 246 millones de años.

se toman esos modelos es que el costo en modelación es muy alto. Es decir, que el modelo analítico ya deja de ser representativo de la Figura 3.1 y por lo tanto deja de ser relevante para este trabajo. El Experimento I es el modelo que desde el punto de vista de modelación habría sido ideal, ya que todos los elementos tienen una justificación ya sea por el análisis exploratorio o por las pláticas con Pablo del Monte. Sin embargo, nuestros resultados nos indican que para inferir con precisión acerca de la extinción con base en la verosimilitud, sería necesario incorporar mayor información. Esto se debe a que la verosimilitud en este modelo es muy plana y por lo tanto la estimación de los parámetros está provista de mucha incertidumbre.

Una verosimilitud plana o no informativa es consecuencia de la presencia de múltiples parámetros de estorbo. Esta noción de verosimilitud plana, es una realidad imposible de eludir en el contexto de querer inferir acerca de extinción con base en el registro fósil. Si ahora agregamos haber identificado la existencia de factores de confusión y proponer maneras de modelarlos tenemos los resultados principales de esta tesis.

Por otro lado el Experimento II muestra una verosimilitud informativa pero no parece ser un modelo adecuado pues las simulaciones indican que el modelo no se ajusta bien a los datos. Por último, el Experimento III pretendía dar la suficiente flexibilidad al modelo del Experimento II para que pudiera ajustar bien a los datos y al mismo tiempo no perder la verosimilitud informativa. Las simulaciones indican que la verosimilitud del Experimento III no es informativa y por lo tanto también haría falta información de otras fuentes para mejorar este aspecto.

En conclusión el modelo parece indicar un buen camino en la estimación de la tasa de extinción ya que de forma manual se muestran casos en los que el ajuste a los datos es muy bueno para diferentes tiempos geológicos. Sugerimos utilizar la expresión 3.4 y los Ingredientes Técnicos Generales como guía para futuras propuestas. Pero más importante, sugerimos fuertemente intentar incluir información de otras fuentes al análisis, por ejemplo, información sobre paleoclimas o aspectos de geología para caracterizar las capas donde fueron encontrados los fósiles. En particular, el esfuerzo de muestreo es un ingrediente que podría ser heterogéneo a lo largo de capas, y muy determinante respecto a la distribución de los datos observados en el registro fósil. El modelo refleja la complejidad que hay detrás de inferir la tasa de extinción a partir del registro fósil y aporta al estudio del problema la comprensión de la necesidad de incluir más fuentes de información al análisis.

3.4. Conclusiones y trabajo futuro

Desde los párrafos introductorios se previó que la inferencia de la tasa de extinción a partir del registro fósil era un objetivo que tenía muchas vertientes que considerarse. El enfoque de modelación estadística que se utilizó en este trabajo, es un enfoque que desde propuestas elementales da información acerca de las dificultades que presenta el problema. Además proporciona un acercamiento al problema que permite introducir información extra cuando se tenga acceso a ella. Tiene la dificultad que depende mucho del conocimiento previo que se tenga del problema, a diferencia de un enfoque de estimación no paramétrica como el que se usó en el Análisis Exploratorio. En resumen, consideramos que la aportación de este trabajo a la estimación de la tasa de extinción a partir del registro fósil es tener una mayor comprensión del problema así como una delineación de los principales problemas que conlleva.

3.4.1. Intuición de la función de riesgo de la evolución

En el capítulo de Análisis Exploratorio se hace mención al trabajo de Nakamura et al. [3] en el cual se utiliza la función de riesgo como herramienta para describir a la tasa de extinción. La función de riesgo asociada a la densidad de longevidades es así un instrumento que permite estudiar el fenómeno biológico de interés. En efecto, por la interpretación que tiene el riesgo, se cuantifica la "probabilidad instantánea de extinción", como función de la longevidad, y su estudio sería portavoz de cambios que han sucedido a lo largo de la historia. Esta noción no debe confundirse con otra función de riesgo, aquella asociada a las longevidades empíricas observadas en el registro fósil, y que ya han sido estudiadas en la Sección 2.2.1 mediante estimadores no paramétricos. Estas segundas longevidades, a diferencia de las primeras, incorporan todos los fenómenos de confusión y sesgo de muestro discutidos en la Sección 3.1. El comentario contenido en la presente sección tiene que ver con el riesgo en el primer sentido, aquel que es de interés biológico directo, y que no es directamente observable sino inferido a través del registro fósil. Se discute qué características generales serían razonables por el contexto y el conocimiento adquirido en esta tesis para esa función de riesgo que radica en el fondo del modelo estadístico considerado. Idealmente, esta tesis proporcionaría un estimador de la densidad de los tiempos de vida a partir de la información proporcionada por el registro fósil. Esto, para posteriormente estudiar al estimador de la función de riesgo de interés biológico correspondiente y de esta forma utilizar las ideas de Nakamura et al. [3]. Sin embargo, ya no se procedió a la construcción de la función de riesgo debido a que el énfasis primordial en esta tesis ha sido examinar el mecanismo de producción física de datos observados ante la acción de diversos factores aleatorios de confusión. El trabajo ayudó a comprender mejor el fenómeno de la evolución y a continuación se describirá de manera intuitiva una posible forma de la función de riesgo de interés biológico.

Para describir a la función de riesgo vamos a distinguir a las especies como tres posibles tipos: especies recientes, especies de vida media y especies longevas. Dentro de las especies de vida media vamos a dividirlas en especies de vida media recientes y las especies de vida media longevas. En la Figura 3.7 ilustramos el comportamiento de la extinción de las especies dependiendo del tiempo que llevan existiendo en la Tierra y a continuación se detallan las ideas que se buscan expresar.

Especies Recientes

Es inmediato suponer que las especies recientes son las más comunes de todas debido al hecho de que la teoría evolutiva tiene como principal consecuencia el cambio constante en las especies mediante la radiación adaptativa. Esto quiere decir que la extinción de una especie no es exclusivamente mediante la aniquilación de sus individuos sino que en muchas ocasiones la extinción de una especie es consecuencia de la evolución. Debido a estos cambios evolutivos suponemos que la tasa de extinción de las especies recientes debe ser más alta que la de las especies que ya llevan más tiempo en la Tierra.

Especies de Vida Media

El motivo por el que se decidió dividir a las especies de vida media en dos grupos fue que se considera que dentro de las especies de vida media se da el momento en que una especie cruza dos umbrales importantes. El primer umbral se refiere a especies que han permanecido suficiente tiempo en la Tierra para que su necesidad de adaptarse ha llegado

un equilibrio, pero aún no lo suficiente para que algún cambio importante en su hábitat (surgimiento de un competidor nuevo o un desastre natural importante por ejemplo) haya ocurrido que los ponga en riesgo. Esto quiere decir que si la Tierra permaneciera constante estas especies difícilmente se extinguirían mediante una radiación adaptativa. El segundo umbral es cuando las especies ya llevan suficiente tiempo en la Tierra y por lo tanto es posible que ocurra un cambio importante en su hábitat en tiempo pequeño. Es por esto que las especies de vida media longevas tienen un mucho mayor riesgo de extinguirse pues las posibilidades de un cambio importante en el hábitat son mayores por el simple hecho de llevar mayor tiempo en la Tierra.

Especies Longevas

Es importante notar que si una especie evoluciona se considera una forma de extinción entonces adaptarse a las condiciones no es una forma de evitar la extinción. Esto implica que las especies deben ser capaces de sobrevivir a los cambios importantes en el hábitat con las mismas características siempre. En cierta medida es un evento fortuito cuando una especie llega a ser especie longeva pues quiere decir que ha sobrepasado cambios importantes sin necesidad de evolucionar. Esto quiere decir que se espera que las especies longevas tengan un riesgo de extinción menor que las especies de vida media longevas, pues ya han sobrepasado cambios importantes a su hábitat. El comportamiento decreciente que se observa en la Figura 3.7 es debido a que mientras más tiempo han permanecido en la Tierra entonces más cambios importantes han sobrepasado y es razonable suponer que estos cambios importantes en cierta medida son similares. Por ejemplo, si la especie sobrevivió a un cambio climático entonces sobrevivirá a otros cambios climáticos y por lo tanto los ciclos en los cambios de las temperaturas de la Tierra no ponen en riesgo su extinción.

Función de Riesgo del Modelo Propuesto

Es natural preguntarnos si la forma de la función de riesgo de los modelos que propusimos tiene un comportamiento al descrito por la Figura 3.7. Para responder esto observemos la Figura 3.8. Se tomó la densidad correspondiente a la ecuación 3.10 con $\theta = (41, 246, 0.025)$, que corresponde a la propuesta con la densidad mezcladora exponencial. Lo que es de recalcar es que usando una densidad con una forma "ideal" según las pláticas con Pablo del Monte y con la intuición que se ha descrito a lo largo de la tesis se obtiene una función de riesgo similar a la de la Figura 3.7. Esto es notable pues no se construyó la función de riesgo correspondiente a los modelos propuestos hasta después de hacer el ejercicio intelectual de imaginar la función de riesgo de interés biológico. Es decir, primero se construyó la Figura 3.7 y después la Figura 3.8. Por lo tanto, la intuición que describimos es congruente con la propuesta del modelo.

Implicaciones de la Concepción de la Función de Riesgo de Interés BIológico

Observando la Figura 3.7 inmediatamente pensamos en que la función de riesgo de interés biológico es genéricamente muy compleja. Esta complejidad necesariamente estará reflejada en la densidad de los tiempos de vida. No obstante, hay que tomar en cuenta que para esta concepción intelectual nos referimos constantemente a los cambios persistentes en el medio ambiente. La idea de densidades condicionales similar al propuesto en la Sección 3.2.1 para abordar los factores de confusión puede ser muy útil para entender la complejidad de los tiempos de vida. Retomemos la variable aleatoria S con densidad $\tilde{f}(s;\tau)$ que modela el tiempo de vida de una especie. La diferencia es que ahora tomaremos una

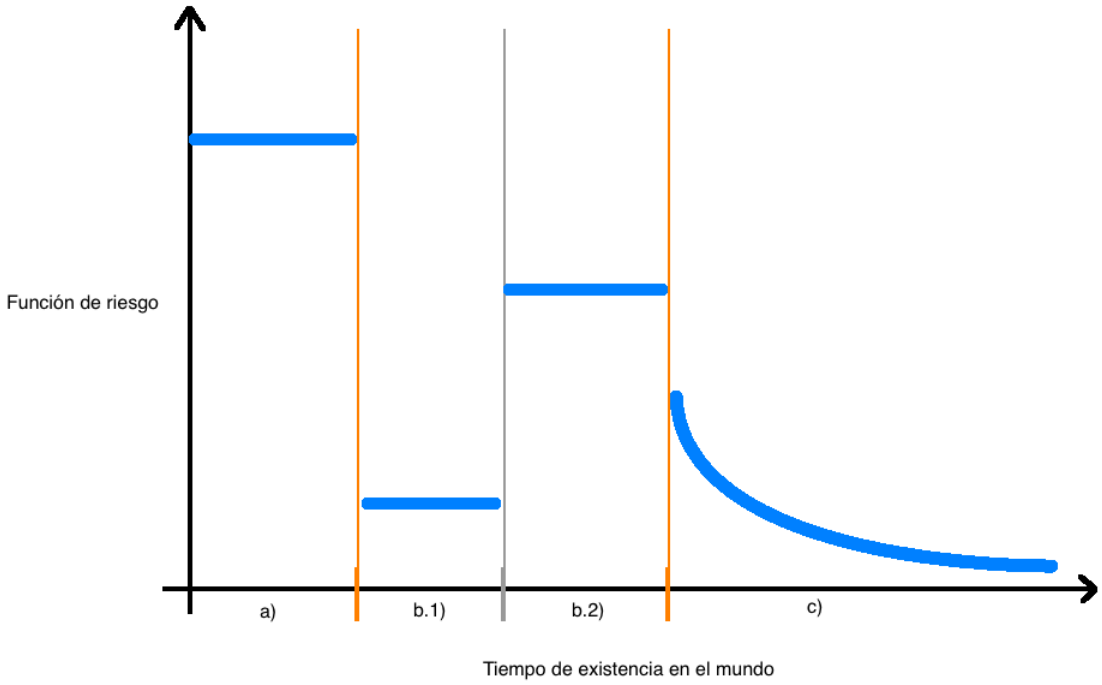

Figura 3.7: Forma esperada de la función de riesgo de la extinción. a) Especies recientes, b.1) Especies de vida media recientes, b.2) Especies de vida media longevas, c) Especies longevas

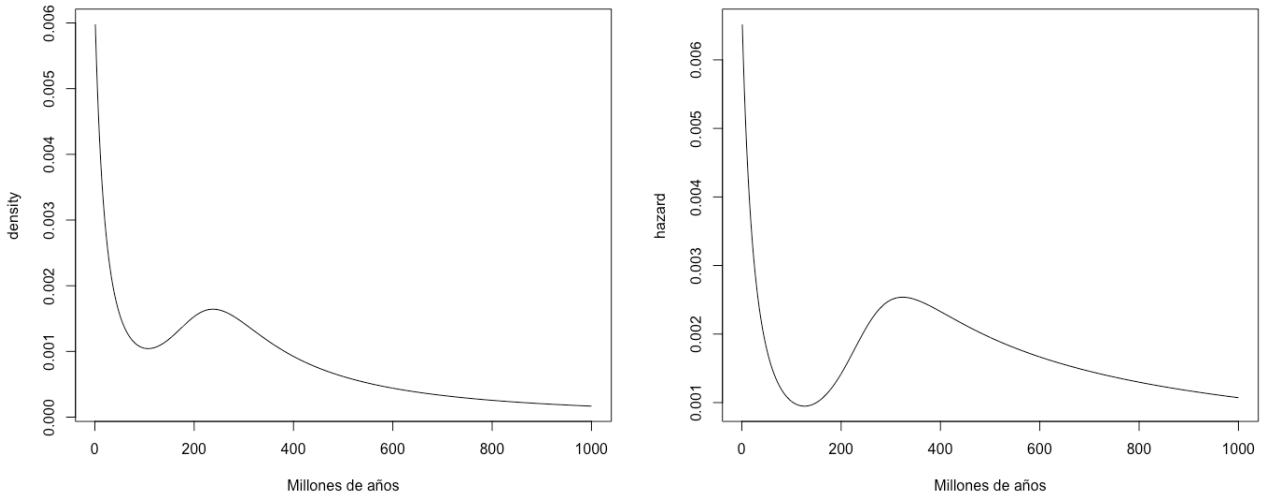

Figura 3.8: Del lado izquierdo de la gráfica se encuentra la densidad de los tiempos de vida con la Propuesta II que se menciona en la Sección 3.2.2 y del lado derecho se grafica su Función de Riesgo correspondiente

variable Bernoulli D^* que modela la probabilidad de que una especie sobreviva (recordemos que sobrevivir significa también no evolucionar a una nueva especie) a un cambio en el medio ambiente sin necesidad de que la especie evolucione y esta variable tiene como probabilidad $p^*(x, \beta^*)$, es decir la probabilidad de éxito (sobrevivir) depende del tiempo de vida. Esto implica que los tiempos de vida de las especies aún pueden tener una densidad con función de riesgo más sencilla pero que va a ser afectada por el inevitable cambio constante del Planeta. Este es otro motivo por el cual es imperante conseguir fuentes de información que hablen de los factores de confusión, ya que la necesitaremos para realizar un modelo de la tasa de extinción aún si tuvieramos datos de los tiempos de vida de las especies. Es decir, la densidad $\tilde{f}(s; \tau)$ no es observable puesto que la única manera de que los cambios en el planeta no afecten a los tiempos de vida de las especies que observamos es cuando la resolución de los datos ya no es en millones de años sino en cientos o miles de años. Sin embargo, para poder tener datos de tiempos de vida es necesario trabajar en la resolución de millones de años.

3.4.2. Conclusiones

1. El registro fósil es la única ventana que tenemos hacia el pasado pero esto no quiere decir que sea una ventana limpia en el sentido que podamos hacer inferencia directamente de la información que nos proporciona. Que sea la única información disponible para indagar acerca de la extinción histórica, no significa que no deba investigarse cuál es la precisión de la inferencia realizada. Por último, tomar en cuenta la presencia de factores de confusión es primordial en la estimación de la tasa de extinción sin importar si se usan los ingredientes técnicos mencionados u otros distintos.

2. La función de detección de probabilidad es un elemento de la modelación que debe ser tomada en cuenta por sí sola como un elemento crucial en la modelación. En ella se puede incluir la información acerca de los factores de confusión y de esta manera aprovechar más la información del registro fósil en la estimación de la tasa de extinción. Además, la noción de datos filtrados por un fenómeno de detección por muestreo es empíricamente compatible con algunas características del registro fósil (en particular, el fenómeno de bimodalidad observado en longevidades para algunos tiempos geológicos).

3. Los modelos aquí presentados nos muestran que la inferencia de la tasa de extinción a partir únicamente del registro fósil es muy complicada. Reconocer esta complejidad nos obliga a considerar la incorporación de información de otras fuentes como un requisito fundamental en el proceso de inferir la tasa de extinción a partir del registro fósil.

4. Hacer una modelación estadística a partir del trabajo interdisciplinario es importante cuando el objetivo de la investigación es entender mejor un fenómeno. Con esto damos reconocimiento de que el modelo estadístico que presentamos tiene un sustento grande en el conocimiento del Dr. Pablo del Monte. Sin la cooperación de mutua este trabajo habría tenido un alcance mucho menor con mucho menor sustento científico en sus propuestas.

3.4.3. Posible trabajo futuro

1. Para poder darle mayor importancia a la función de detección de probabilidad se deben conseguir datos de otras fuentes. Con esto, la información que aporta el registro fósil a la estimación de la tasa de extinción se diluye menos en los factores de confusión. Posiblemente con la incorporación de otra información los elementos técnicos específicos de la expresión 3.1 no se deban modificar y sólo se deba enfocar el trabajo en determinar cómo incluir las nuevas fuentes de información dentro de la función de probabilidad de detección.

2. Incluir en la modelación el hecho de que los datos son censurados. Este es un elemento que debe ser incluido eventualmente pues es una característica de los datos que no debe ser ignorada en el análisis final. Sin embargo, va a generar mayor incertidumbre y por lo tanto se recomienda no hacer este avance hasta haber conseguido información de la probabilidad de detección.

3. El esfuerzo de muestreo debe ser uno de los factores de confusión que se deben intentar resolver primero. Esto es debido a que afecta fuertemente la distribución de los datos. Además, se trata de un factor de confusión muy difícil de incluir en la función de probabilidad de detección sin tener algún proxy acerca de su comportamiento.

4. Sugerir densidades mezcladoras que simplifiquen el modelo y tengan una génesis adecuada. Las propuestas que se enuncian en este trabajo no exhaustivas y por lo tanto se reconoce la posibilidad de que se encuentren elementos técnicos específicos que permitan hacer una estimación de la tasa de extinción usando exclusivamente el registro fósil. Sin embargo, debe tomarse en cuenta que este camino puede demandar mucho tiempo y esfuerzo sin resultados diferentes a los aquí presentados. Se recomienda tomar esta dirección únicamente si se tienen motivos específicos para hacerlo y no de forma exploratoria.

5. Proponer otras funciones de probabilidad de detección que reflejen el conocimiento que se mencionó pero que faciliten las expresiones analíticas. Aunque no se muestra en los resultados de este trabajo, debemos aclarar que se exploró el caso en que la función de probabilidad de detección era escalonada sin éxito. Sin embargo, alentamos a que se retome el caso pues una manera de mejorar la forma de incorporar la información de otras fuentes.

Bibliografía

[1] *http://www.biodiversidad.gob.mx/especies/extincion.html*

[2] D. Schluter, *The ecology of adaptive radiation*, Oxford University Press, pp. 10-11, 2000.

[3] M. Nakamura et al., *Statistical inference for extinction rates based on last sightings*, Journal of Theoretical Biology, Vol. 333, pp. 166-173, 2013.

[4] J. J. Sepkoski Jr., *Biodiversity: past, present and future*, Journal of Paleontology, Vol. 71, No. 4, pp. 533-539, 1997.

[5] M. Novacek *The biodiversity crisis: losing what counts*, The New Press, 2001.

[6] D. Barnosky et al., *Has the Earth's sixth mass extintion already arrived?*, Nature, Vol. 471, pp. 51-57, 2011.

[7] J. Alroy, *Phanerozoic trends in the global diversity of marine invertebrates*, Science, Vol. 321, pp. 97-100, 2008.

[8] D. Huang, E. Goldberg, K. Roy, *Fossils, phylogenies, and the challenge of preserving evolutionary history in the face of anthropogenic extinctions*, PNAS, Vol. 112(16), pp. 4909-4914, 2015.

[9] C. R. Marshall, *Marine biodiversity dynamics over deep time*, Science, Vol. 139, pp. 1156-1157, 2010.

[10] D. Burney y T. Flannery, *Fifty millennia of catastrophic extinctions after human contact*, Vol. 20(7), pp. 395-401, 2005.

[11] P. Harnik et al., *Extinctions in ancient and modern seas*, Vol. 27(11), pp. 608-617, 2012.

[12] J. Nelder y R. Mead, *A simplex algorithm for function minimization*, Computer Journal, Vol. 7, pp. 308-313, 1965.

[13] V. Proenca, *Comparing extinction rates: past, present, and future*, Encyclopedia of Biodiversity, Vol. 2, pp. 167-176, 2013.

[14] A. Solow y W. Smith, *Missing and presumed lost: extinction in the ocean and its inference*, ICES Journal of Marine Science, Vol. 69(1), pp. 89–94, 2012

[15] t. Bouezmarni, A. El Ghouch y M. Mesfioui, *Gamma kernel estimators for density and hazard rate of right-censored data*, Journal of Probability and Statistics, Vol. 2011, Article ID 937574, 16 pages, 2011

[16] *http://www.bbc.co.uk/nature/extinction_events*

[17] A. El-Shaarawi, W. Piegorsch (eds.), *Enciclopedia of Environmetrics*. John Wiley & Sons, Ltd, 2002.

[18] J. Àitchison, J. Brown, *The Lognormal Distribution*. Cambridge University Press, 1963.